

7461901-quoting-the-touch-does-not-pay-its-adverse-selection-a-true-aggressor-signed-ent

AVTORJI: Daniel Gatto

POVZETEK:

Raziskava obravnava, kako se odzivajo trgovci pri zagonu z uporabo najboljšega nabira in najboljšega ponudnika (best bid/offer) v trgovini kriptovalut. Natančno merila, ki jih uporablja avtor, razdelijo vsak preplon na polpribližek, ki ga je trgovec zaznamal pri trgovini, in potezo srednje cene v času merila. Rezultati kažejo, da je vse tri kriptovalne trgovine (Bybit, Binance USD-M in Hyperliquid) negativni preplon pri zagonu z najboljšim nabiro in najboljšim ponudnikom, pri čemer je negativnost preplonov na vseh trgovinah preden je dodana naročniplačila. To pomeni, da je trgovska selekcija (adverse selection) večja od polpribližka, ki ga je trgovec zaznamal pri trgovini.

METODA:

Avtor uporablja merila, ki razdelijo vsak preplon na polpribližek, ki ga je trgovec zaznamal pri trgovini, in potezo srednje cene v času merila. To merilo je uporabljal tudi na trgovinah CME futures, kjer je merilo negativnost preplonov pri zagonu z najboljšim nabiro in najboljšim ponudnikom. Merila so izvedena na podlagi 9,983 kriptovalnih dnevnikov in 125,828,922 simuliranih preplonov. Avtor tudi preverja, kako so različne merila uporabljena na različnih trgovinah in kako se njihove rezultate ujemajo s merili, ki jih uporablja trgovina.

KLJUČNE UGOTOVITVE:

- Vse tri kriptovalne trgovine (Bybit, Binance USD-M in Hyperliquid) imajo negativen preplon pri zagonu z najboljšim nabiro in najboljšim ponudnikom.
- Negativnost preplonov je preden je dodana naročniplačila.
- Trgovska selekcija (adverse selection) je večja od polpribližka, ki ga je trgovec zaznamal pri trgovini.

POMEN ZA TRGOVANJE:

Raziskava je pomembna za trgovce, ker ukazuje na to, da je trgovska selekcija večja od polpribližka, ki ga je trgovec zaznamal pri trgovini. To pomeni, da je trgovec pri zagonu z najboljšim nabiro in najboljšim ponudnikom v več kot polpribližek zmanjšen. Trgovci bi morali upoštevati to, ko izvirajo informacije o ceni, da bi lahko bolje upravljali svoje riske in prihodki.

Quoting the Touch Does Not Pay Its Adverse Selection

A True-Aggressor-Signed Entry-Markout Decomposition for a
Last-in-Queue Quoter

Across Bybit, Binance USD-M and Hyperliquid

Daniel Gatto^{*1}

¹Universidade Paulista (UNIP)

August 2026

Abstract

A passive quote earns the half-spread and pays adverse selection to whoever consumes it, the Glosten–Milgrom mechanism. On CME futures the two legs offset, +0.654 bp against −0.661 bp, though queue rule confounds that contrast: the companion paper’s order sits at the front of the remaining queue, this one at the tail. Across three crypto perpetual venues, 9,983 coin-days and 125,828,922 simulated fills, a queue-at-the-touch quoter’s pooled 10-second pre-fee entry markout is negative at each: Bybit −0.8387 bp, Binance USD-M −0.3912 bp, Hyperliquid −0.4575 bp, with two-way month-and-coin intervals excluding zero at all three. Because these exchanges publish the aggressor flag most studies must infer, its cost is measurable: Lee–Ready and the tick rule agree with the exchange on 89% to 98% of fills, preserve the sign, move the magnitude by less than a quarter, and do not rank by agreement rate. The quantity is one-sided: no exit and no exit cost, so it cannot say whether these market makers earn or lose money. The novelty is not the negative sign but the signed decomposition and the priced cost of trade-sign inference. Two boundaries are reported, not smoothed over: the reproducible chain begins at frozen derived artifacts, not raw archives, and two shipped extractions of the same 500 Hyperliquid coin-days differ twofold in fill count and by 0.2390 bp in net, where Bybit’s agree exactly.

Keywords: market making, market microstructure, adverse selection, entry markout, trade-sign classification, cluster-robust inference, multiway clustering, perpetual futures, decentralized exchange, cryptocurrency, reproducibility.

JEL classification: G14, G12, C58, C12, D47.

^{*}Correspondence: daniel@daru.finance. Reproducibility package: <https://github.com/DaruFinance/crypto-adverse-selection>.

Contents

1	Introduction	3
2	Related Work and the Gap	5
3	Data	7
4	Measurement Rule	9
5	The Central Result	13
5.1	Per Coin, Per Venue	15
6	Trade-Sign Counterfactual	17
7	Inference	20
7.1	Months alone	20
7.2	Interval calibration	21
7.3	Withheld verdicts	22
8	Venue Comparison	23
8.1	Lead-lag	25
9	Conditional Decomposition	25
10	Depth and Rebate Frontier	26
11	The Quoting Arm	28
12	What the Measurements Decide	30
13	Reproducibility	31
14	Discussion and Limitations	33
15	Conclusion	36
A	Per-Coin	37
B	Panel Lineage	38
C	Figure Provenance	39
D	Results Inventory	39
	Data and Code Availability	40
	References	41

1 Introduction

Market making posts a bid and an offer and waits, earning the half-spread for standing ready to trade while paying when an arriving counterparty holds information the price does not yet reflect, which is the adverse-selection cost of Glosten and Milgrom [20]. The companion paper [18] measures this trade by trade on CME depth-10 futures with true aggressor flags: a captured half-spread of +0.654 bp against a post-trade drift of -0.661 bp, with net values indistinguishable from zero at each of thirteen roots. Posing the same question to crypto perpetual order books inverts the answer. Adverse selection does not offset the touch but over-consumes it, and the simulated touch quoter’s entry markout is negative on all three exchanges before any fee is charged. The inference work is the other half of the contribution, because the crypto panel is hostile terrain for cluster inference. Fill weight concentrates in a few coins, calendar months are few and unevenly sampled, and one exchange’s month distribution is right-skewed at +2.02, which lifts the sign-flip test’s size to 0.124 against a nominal 0.05. Section 7 therefore reports which statements survive rather than assuming any do. The one-sentence takeaway is that the sign of the net entry markout survives whereas almost everything else in its neighbourhood does not.

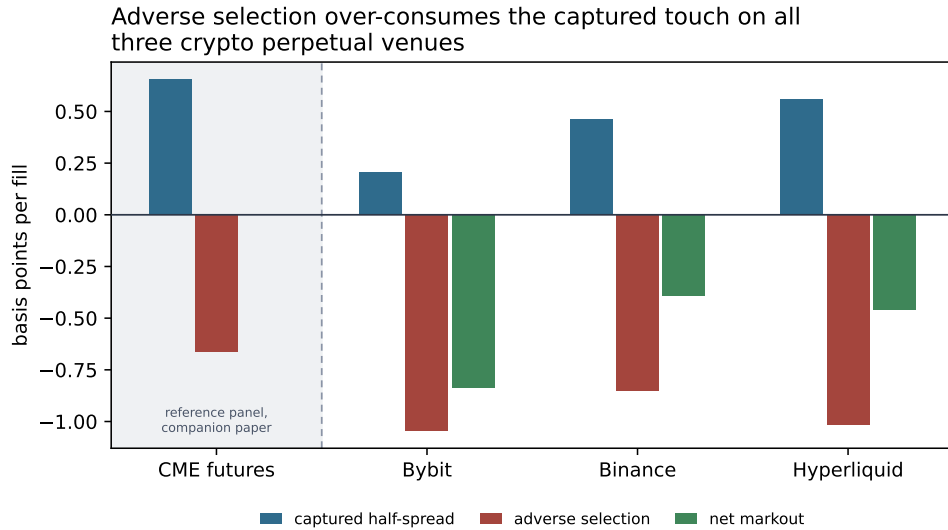


Figure 1: For simulated queue-tail best-bid/offer quoters, the captured half-spread is smaller in magnitude than the post-trade adverse drift on all three crypto exchanges, so the pooled 10-second pre-fee entry markout is negative: $-0.8387 / -0.3912 / -0.4575$ bp. The three crypto panels cover different calendar intervals and different coin sets, so these bars do not constitute an exchange ranking. The shaded group at far left is not a fourth measurement; it reproduces published CME companion-paper values [18] from a different asset class, period and data source, and measured for a quoter placed at the FRONT of the remaining queue rather than the tail, shown as a reference point for the offsetting case, not a comparable control; only its two legs are drawn, since the point of the reference group is that they offset and the net they imply, -0.007 bp, is indistinguishable from zero at the precision published; the queue rule alone moves the capture leg, so the contrast is not attributable to the asset class. Source: crypto groups from `reproduce/decomposition_by_venue.json`; reference group from [18].

The decomposition that delivers this result is built on a specific object: the entry markout, which splits a fill into the half-spread captured relative to the prevailing mid at the time of trade plus the drift of that mid over the markout horizon, with no closing

trade and no exit cost. A negative entry markout therefore does not say the position loses money, because the exit is unmeasured and could carry revenue of its own. It does not say the round trip is whole either: a passive exit at the touch is this same measurement pointed the other way, and it is negative on all three venues. The object measured is one particular simulated quoter, not “the market maker”: one unit posted at the best bid and one at the best offer, re-quoted at every snapshot, queued behind all visible size at the price, and filled only when volume exhausts that queue. Consequently, every claim in this paper is a claim about that quoter. Two neighbouring quantities do not survive the way the sign does, and they are rejected where they appear rather than here: the adverse-selection-to-capture ratio (Section 5) and the split between the two legs (Section 4).

Measuring at ten seconds and pooling across the panel, the simulated queue-tail touch quoter earns a negative fee-free entry markout on all three venues, and the contributions of the paper are six. (a) The decomposition. An entry-markout decomposition with true aggressor flags for the simulated queue-tail touch quoter on three crypto perpetual exchanges, two centralized and one on-chain CLOB, built on 9,983 coin-days and 125,828,922 simulated fills. The pooled 10-second fee-free net is negative on all three, the two-way interval over months and coins excludes zero. The verdict survives all four units on Bybit and Binance and the two 10-second units on Hyperliquid, whose 60-second units are withheld, and leave-one-coin-out. This inverts the near-exact cancellation that the same decomposition produces on CME futures [18]. (b) Inference sized to the panel. Two-way clustering over months and coins rather than months alone; a Kish effective cluster count; withheld verdicts where the interval is not worth reading; and calibration of four interval methods through the identical pipeline, with the two conditions that decide the month-only-versus-two-way choice re-run in this paper’s own clustering, to locate the region where they fail together rather than cherry-picking the one that passes (Section 7.2). (c) Trade-sign counterfactual. Measuring classification error against a perfect benchmark has precedent, in Odders-White [27] on equities and Andersen and Bondarenko [7] on order-flow toxicity, but the new step is to push that error through the downstream quantity of the maker-side entry markout on crypto perpetuals. Re-running the whole simulation under Lee–Ready and under the plain tick rule, each of which takes its own fill set, on all three exchanges, the classifiers agree with the exchange flag on 89% to 98% of fills, change the sign nowhere, and move magnitudes by 0.6% to 20.4%, with the direction of the move differing between centralized and on-chain venues. Agreement rates do not rank this error across venues, so they yield neither a constant correction nor an agreement-rate threshold. (d) Negative and scope-limiting results reported as findings. A descriptive, sampling-dependent venue comparison; a lead-lag relation inseparable from publication-rhythm differences; a conditional decomposition that is a low-power null; a depth-and-rebate frontier whose dominant sensitivity falls outside its own interval; and a quoting arm whose apparent advantage over the touch reverses above a rebate the panel locates, with which side a given maker falls on decided by that maker’s own fee schedule. (e) Timestamp-collision measurement. The share of fills sharing a timestamp with an order-book update is Bybit 0.3%, Hyperliquid 8% to 10%, and 52.7% on millisecond-stamped Binance USD-M; switching the quote read from strictly earlier than to earlier than or equal to the fill drops the quote rule from 0.996 to 0.620 on one heavy Binance day. This is a property of the data source’s clock, not of the classifier, and it determines whether the three venues are comparable; these particular numbers are diagnostics measured outside the reproducible chain and are labelled as such wherever they appear. (f) Reproducibility boundary. All ten aggregate panels rebuild byte-for-byte

from hash-verified frozen artifacts, and every shipped result file is hash-checked against the same manifest. Whether a result file also rebuilds is a second question with a more uneven answer: the shipped-versus-rebuilt comparison covers every leaf of three of them and as little as 68 of 6,575 on the cross-venue lead-lag, so a clean comparison speaks for the numbers it reaches rather than for the whole file. The raw archives sit explicitly outside that chain.

Because the claim is that the simulated touch quoter’s fee-free entry markout is negative on these exchanges, the falsification condition is concrete: a unit in which adverse selection is consistently below the captured half-spread, with the net excluding zero on the positive side. Seven of Hyperliquid’s twenty coins carry positive point estimates at 10 seconds, led by NEAR at +0.4029 bp, though five of those seven carry a withheld verdict and neither of the two tested excludes zero, so no cell in these panels yet meets the condition. Section 14 states which tests such a unit must survive, and the per-coin files shipped with the paper are organized so that such units are immediately visible.

2 Related Work and the Gap

Glosten and Milgrom [20] model the bid–ask spread as compensation for adverse selection. Moallemi and Yuan [26] price the value of queue position against the drift a resting order suffers after execution, whereas this decomposition holds queue position fixed at the tail and measures the drift a fill at that position carries, so the two are adjacent objects. The direct predecessor is the CME companion paper [18], which applies the identical split to depth-10 futures stamped with true aggressor flags and finds capture approximately equal to adverse selection, with net values indistinguishable from zero at every root. The contact point is the inversion: the same decomposition, applied to exchanges with different tick structures, fee schedules, and matching architectures, places the adverse drift *above* the captured half-spread rather than alongside it.

That inversion separates this paper from the AMM line of work, and the distinction is stated explicitly rather than assumed, because a reader encountering both titles would otherwise treat them as overlapping. The second companion paper [19] studies DEX markets, reconstructing on-chain OHLCV to examine liquidity provision in AMM pools and order flow in DEX-native tokens. AMMs have no order book, no queue, and no discrete quote to be consumed; their LPs hold a continuous curve and pay arbitrageurs the loss-versus-rebalancing that Milionis, Moallemi, Roughgarden, and Zhang [25] price. Hyperliquid, by contrast, is an on-chain central limit order book with discrete price levels, visible queues, and exchange-published aggressor flags. The object here is a resting limit order, and the cost is the same Glosten–Milgrom takedown cost measured on two centralized exchanges. The two papers share a role-level primitive and nothing else in their measurement.

The empirical microstructure literature on these three venues differs in what it measures. Albers, Cucuringu, Howison, and Shestopaloff [1] run live experiments with millions of orders on Bybit and Binance, modeling how latency degrades execution relative to the order book snapshot at submission. That work establishes a taker-side execution quality result for the two centralized venues. This paper instead measures the markout of passive quotes. Barone and Lillo [9] study adverse selection on Hyperliquid directly, reconstructing 4.3 million hidden metaorders from address-level data and comparing them against 465,000 visible protocol-native TWAP executions and testing whether publicly

disclosed intent lowers execution cost. Their object is parent-order market impact from the taker side, whereas the drift borne by a resting passive quote after execution is the object here. Same primitive, opposite sides of the trade, and neither substitutes for nor duplicates the other, which is why the title must state what is measured. On the data side, Albers and co-authors document Hyperliquid’s level-4 data source and its three data streams [4].

Because the three crypto venues used here publish an aggressor flag on every message, this paper can measure trade direction instead of assuming it, which the equity microstructure literature cannot do. That literature infers the aggressor from prices and volumes because exchanges do not publish an aggressor field on message-level feeds such as the US consolidated tape, Nasdaq ITCH, LSE MITCH, and Euronext Optiq, whereas CME MDP 3.0 and Eurex T7 do, as does the data source behind the CME companion paper [18]; inference is needed where the resting side cannot be reconstructed from the feed, as when executions reference hidden or otherwise opaque orders. Lee and Ready [23] provide the standard substitute, and its errors concentrate in high-volatility and high-frequency regimes and where hidden orders are active [15, 17]. None of the headline results here involve a tick rule, quote rule, or volume classifier, since the true aggressor flag is observed directly; the classifiers enter only as the counterfactual of Section 6. Having the true direction makes classifier error measurable rather than assumed, which is the content of Section 6, and that comparison requires a ground-truth direction and is therefore unavailable to the literature it addresses.

A passive quote that is taken is a trade, and the inventory literature prices how a market maker should shift quotes as inventory accumulates. Avellaneda and Stoikov [8] give the classic model of inventory-skewed quoting, extended by Guéant, Lehalle, and Fernandez-Tapia [21] and by Cartea, Jaimungal, and Penalva [14]. Section 11 reports a restricted quoting arm on a single venue as one tested rule, not a test of inventory-aware market making as a whole, because its own diagnostics show the tested object behaves as a depth rule rather than a faithful implementation of the model.

Across the panel, inference is where the work is. The panel is coin-day with two natural dependence dimensions, calendar month and coin, which are not nested. Cameron, Gelbach, and Miller [13] supply the multiway estimator used for the headline intervals. Their earlier bootstrap work [12] and MacKinnon and Webb [24] address the few-clusters problem and the sensitivity of cluster inference to unequal cluster sizes. Every panel here sits in that regime. Petersen [28] is the finance reference point for how much standard errors move with the choice of clustering dimension, and Kish [22] gives the effective sample size that determines whether a clustered interval is worth interpreting. The contribution is not a new estimator but a practice: such results are usually reported under a single clustering choice with alternatives left uncomputed, whereas here the month-only and two-way month-and-coin intervals are computed alongside the four interval methods of the calibration, all disagreements are reported, and Section 7.2 locates the region where the family fails together.

That practice separates this paper from the nearest prior work on the passive side. Albers et al. [2] price the trade-off between fill probability and post-fill returns faced by a passive quoter, which is the same economics. DeLise [16] models the negative drift attached to limit order executions, and Tiniç, Şensoy, Akyildirim, and Corbet [29] measure adverse selection in cryptocurrency markets through spread decomposition rather than by simulating a quoter’s markout. Four differences follow: (i) three venues under one identical queue simulator rather than one; (ii) 9,983 coin-days and 125,828,922 simulated fills rather than a single experiment; (iii) every fill is direction-stamped by the exchange’s

own aggressor flag, with the classifier counterfactual measured rather than assumed; and (iv) a formal split of the captured half-spread from the post-fill drift. Those works precede this paper in the specific respect of modeling the passive quoter’s trade-off, and this paper extends rather than replaces them.

Hyperliquid maker-side markout already has a literature, and the closest measurement in it is worth stating precisely, because it is the nearest external check this paper’s central number has. Albers and coauthors read rejected orders as return predictors [3], and, from the complete message flow of the BTC-USD perpetual over July 2025, identify a cohort of eight traders that wins first-in-line queue position by posting inside the spread ahead of the widening [5]. In the course of that they measure the same estimand this paper measures, the signed midprice change at a horizon relative to the fill price taken from the maker’s side, on real fills rather than simulated ones. Two of their figures matter here. The average maker fill on the venue loses about 0.18 bp at five seconds net of the maximum 0.3 bp rebate, so roughly -0.48 bp before it; the queue-priority cohort’s own fills earn $+0.32$ bp on the same basis. The first sits close to the -0.4575 bp this paper reports for Hyperliquid at ten seconds, and to the -0.5832 bp it reports for Hyperliquid BTC, which is the instrument they measure. The agreement is directional and not a replication: their horizon is five seconds against ten, their weighting is by volume against this paper’s by fill, their window is one month against eleven, and their object is every maker fill on the venue against one simulated quoter. What it does establish is that the level this paper reports for that venue is not an artefact of simulating the quoter, since a measurement on observed fills lands in the same place. The second figure is the queue-position gap, and Section 14 returns to it. Barone and Lillo [9] measure adverse selection from the taker side. None reports the passive quoter’s own entry markout decomposed into its two legs, which is what Section 5 delivers and Section 6 then reruns under every classifier.

The decomposition is new in a specific sense. It is not new that passive quoting at the touch carries a negative entry markout before rebates; that is informal consensus in practitioner tooling and was measured on a single venue by Albers and coauthors [2], so the sign is not the surprise. What is new is the first three-venue, ground-truth-direction decomposition of it, applying one quoting rule across two centralized exchanges and an on-chain CLOB, and measuring on the same fills the cost of the aggressor flag itself.

3 Data

Across the three venues, the panels are built from exchange-native data feeds, one per venue, and every row of each panel is a coin-day carrying that coin’s fill count and the fill-weighted decomposition components for that date. The measurement rule requires the visible resting size at the touch, so each feed is at least top-of-book with size. However, the shipped panels record only the coin-day aggregates. Consequently, the snapshot construction, the tick sizes, the quoted spreads and the number of quoting opportunities behind the headline panels are not in the package and are not reported here; tick sizes are venue specifications rather than panel contents, and the two used in Section 5 are quoted there from Bybit’s published contract terms. Table 1 summarises the panels; the fills counted there are the simulated fills of the reference touch quoter defined in Section 4, not every passive fill on the venue, and the maker fee column reproduces the venue’s published schedule for reference only (all headline numbers are pre-fee).

Table 1: Panel coverage

venue	coin-days	coins	dates	months	fills	window	fee (bp)
Bybit perpetuals	3,618	13	295	29	45,476,018	2023-04 to 2025-08	2.0
Binance USD-M	4,785	15	319	11	78,408,624	2023-05 to 2024-03	2.0
Hyperliquid	1,580	20	79	11	1,944,280	2025-07 to 2026-05	1.5
Total	9,983				125,828,922		

Each row aggregates a coin-day panel for one venue. Fills are simulated executions against the Section 4 reference-quote maker, not every passive fill on that venue. The maker-fee column reproduces each venue’s fee schedule for reference only; all headline figures are pre-fee. Source: `reproduce/decomposition_by_venue.json`

All three venues publish an exchange-side aggressor flag on every fill, which is the raw material this decomposition requires and which futures or equity research usually has to infer. Beyond that commonality, the feeds are not equivalent, and four differences matter for what can be read from the pooled numbers.

First, the panels cover different periods, and only two of the three pairs overlap at all. Binance and Hyperliquid share no dates; the Bybit and Hyperliquid main panels overlap on twelve dates and 156 coin-days; and the Binance window lies entirely inside Bybit’s, so those two share 107 dates and 1,284 same-date coin-days on 12 common coins. The re-quoted panel behind Section 8 is a separate draw with 25 common dates. Venue comparisons read directly from the pooled Table 3 numbers are therefore confounded with the periods each panel covers, so Section 8 runs only on common coin-days and is still reported as descriptive.

Second, sampling density differs, and on one venue the schedule is not a stride at all. Binance covers every calendar day from 2023-05-16 to 2024-03-29, a window of exactly 319 days. Bybit covers 295 dates on a uniform three-day stride, roughly ten per month in every one of its 29 months. Hyperliquid’s 79 dates average one day in four, but that average describes no part of the schedule: 26 of them are consecutive, running 2025-07-27 to 2025-08-21, and the remaining 53 sit on a five-day stride from 2025-08-26 to 2026-05-28, from which three dates are skipped. The block puts 23 of the 79 dates and 33.7% of that venue’s fills inside 2025-08 alone, against four to six dates in each of the other ten months. Every property this paper later reports of the Hyperliquid month dimension follows from that shape rather than from the venue: 6.30 effective months of 11 on the Kish count, the cross-month skew of +2.02 that Section 7.2 feeds to the sign-flip calibration, and the 11 of 20 per-coin verdicts the effective-month floor withholds, among them ETH at 4.61 effective months despite an interval of $[-0.7962, -0.6058]$. So the month counts of 29, 11 and 11 describe clusters weighted very differently from one another, and the Hyperliquid eleven are the least evenly weighted of the three. The Bybit stride also explains why the twelve Bybit/Hyperliquid overlap dates are an artefact of the two sampling schedules rather than the true calendar overlap, which is wider.

Third, the coin sets differ: 13, 15, and 20 coins, with 12 coins in common. Restricting to those 12 raises the adverse-selection-to-capture ratio on the two venues where it is lowest: Hyperliquid from 1.82 to 2.21 and Binance from 1.85 to 2.06, with Bybit essentially unchanged at 5.06. No interval on this ratio is reported anywhere in this paper.

Fourth, fill counts are not comparable across analyses within a venue, let alone across venues. The three Bybit panels count different things: the same coin-day carries 9,765 fills in the main decomposition, 5,743 in the quoting-arm touch leg, and 2,789 at the depth-frontier touch, because each analysis runs its own quoting rule. The depth frontier is buy-side only, which halves the count to about 4,882, and the differing subsample and

rung grid account for the remaining factor of roughly 1.75 against the 2,789. A fill count in this paper indexes only the analysis in which it appears and stands for nothing broader.

The addressable universe is the set of crypto-perpetual venues that publish a true aggressor flag on every fill and also maintain a retrievable order book or top-of-book archive, over each venue’s historical and liquid coin list. This paper measures three of them, two centralised and one on-chain, a deliberate spread across matching architectures, not a census. OKX, Coinbase, Deribit and other centralized books are untested and are explicitly named as untested units. Within each venue, the panel is a subset in both dimensions at once, coins and dates: 13/15/20 coins; Binance covers every calendar day, Bybit a uniform third of them, and Hyperliquid a consecutive block followed by a five-day stride. Coverage cannot be quantified from the accompanying deliverables, for a reason stated plainly: the coin lists and sampling strides are fixed by upstream measurement procedures that sit outside the reproducibility package, so the package records which coin-days entered the panel but not the admission rule. Any claim that this is n percent of a venue’s liquidity universe is unsupported by the accompanying deliverables, and this paper makes no such claim.

Every coin in every panel is still trading on the panel’s last sampled date, which makes each panel a survivor set by construction. Under the survivor-set construction, the panels are still not point-in-time. The cost of that is measured in Section 14 and in the tenth accompanying panel: 3,451 coin-days on 14 perpetuals that were delisted within the window, measured under the identical rule. For Bybit, delisted names underperform survivors by 0.8483 bp on matched dates, and omitting them moves the pooled number by roughly 0.0720 bp in the direction of making the measured net value less negative than the truth. For Binance USD-M, the same test returns a withheld verdict on the four delistings inside its 319-day window, four coins being below the effective-cluster floor. For Hyperliquid, the test cannot be run at all. The positioning statement is that the bound is a bound and not a correction: it says how far the estimate could be off, while the panel remains survivor-biased.

4 Measurement Rule

That inversion, a passive quote that is crossed by the market and pays away more in post-fill drift than it earns in captured half-spread, is the central object of this paper, and the measurement rule that produces it must be stated exactly before any result can be read. The rule is deliberately conservative: a simulated maker posts one unit at the best bid and one at the best ask on every snapshot, and only when both sides exist, re-quoting at every snapshot so that each order lives for exactly one snapshot interval with no priority carried over. The quoter re-quotes no more often than every 100 ms, so the effective interval is whichever is longer, that floor or the venue’s own publication cadence: about 100 ms on Bybit and about 541 ms on Hyperliquid, as Section 8.1 reports. Table 2 corroborates this only in part, and the part it fails is informative. The 541 ms figure is a median snapshot spacing rather than a fixed grid, so a re-quote floor below it still binds on the short gaps in the distribution: the Hyperliquid fill count moves by 0.1% from the 100 ms rung to the 250 ms rung, which is the behaviour a cadence well above both would produce, but then falls a further 4.4% at the 500 ms rung, which is still below 541 ms and should have been non-binding under a fixed grid. Every Bybit rung binds throughout. Both readings come from the re-quote panel, which Section 8 shows does not agree with the Hyperliquid

main panel, so they speak for the cadence of the re-quote extraction rather than for both. Binance’s snapshot cadence is not recorded in the shipped panels and no rung sweep was run for that venue. The shipped sign-rule file carries 29,949 coin-day rows, three rules times the full 9,983 coin-day headline set, and its exchange-flag rows reproduce Table 3 exactly. The trade-sign counterfactual of Section 6 runs the reference quoter at that same 100 ms re-quote ceiling, which is what the shipped counterfactual file records. Table 2’s rung sweep is a separate 25-date slice measured under its own gap grid, so its absolute levels are not comparable with the headline and only its rung-to-rung direction is. When an order is placed it joins the back of the queue behind all visible resting size at that price. It fills only when an aggressor arrives on the opposite side, the trade price equals the quoted price, and the trade size exceeds the queue still ahead of it. In that case, the fill quantity is the overflow, capped by the amount still resting. An order posted at the final snapshot has no subsequent snapshot to cancel it. It therefore persists to the end of the tape, which in a trading day of hundreds of thousands of snapshots concerns exactly one of them. The mid-price is read by a staircase rule: the last snapshot strictly before the time in question, ties broken on receive time, with no interpolation. Taking the last snapshot at or before the time would read the post-fill book wherever a fill and the update it triggers share a timestamp, which is 52.7% of fills on millisecond-stamped Binance USD-M.

This sits on the pessimistic side, and that matters for how to read the result. Queueing behind all displayed size and filling only on overflow is the least favorable queue assumption short of never filling at all, so a real maker holding genuine time priority at the same price would capture more of the half-spread and be adversely selected less often; a negative entry markout under this rule is therefore not evidence that a maker with priority would also be negative. That is no longer only an argument from mechanism. On Hyperliquid, Albers and coauthors [5] measure the same estimand on observed fills and put the average maker fill at roughly -0.48 bp before rebate while a cohort that wins first-in-line position earns $+0.32$ bp, so on that venue the two ends of the queue sit on opposite sides of zero and about 0.8 bp apart. Their figures are not a calibration of this rule, for the reasons Section 2 lists, but they fix the direction and the rough size of what the queue assumption here gives up. The companion CME paper [18] takes the opposite assumption, placing the order at the front of the remaining queue, and states its results as single-sided bounds in the other direction. The two papers sit on opposite sides of the queue assumption, front and tail, but neither this paper nor the pair of them is offered as a bound on the true value, because the two differ in asset class, period and data source as well as in queue rule.

The decomposition is defined by three equations, which fix the vocabulary for every table that follows. Let s be the maker’s side ($+1$ when the maker ends up long, meaning the aggressor sold), m_0 the mid-price at the fill, m_τ the mid-price one horizon later, and f the fill price, all expressed as ratios of m_0 and reported throughout in basis points. Then capture $= s(m_0 - f)/m_0$, adverse $= s(m_\tau - m_0)/m_0$, and net entry markout $= \text{capture} + \text{adverse} = s(m_\tau - f)/m_0$. The horizon $\tau \in \{10s, 60s\}$, with the headline specification at 10 seconds. The direction is taken from the venue’s true aggressor flag, inverted, because the tape records the aggressor’s side whereas the equations take the maker’s.

Holding the re-quote interval fixed, the split between the two legs is sensitive to the reference mid in a way that the net is not, and this distinction carries through every later section. The reference mid m_0 enters (1) and (2) with opposite signs, so for a given fill it cancels in the numerator of the net, leaving the net unchanged up to a positive rescaling and only moves the boundary between the two legs. Pricing against a staler reference

books the pre-fill drift into the adverse term and inflates measured capture by the same amount, so the split is an artifact of quoting cadence whereas the net, for a fixed fill set, is not. But changing how often the maker re-quotes is not merely a restatement of this split, because it also changes which fills occur at all. Two channels are visible in Table 2: a staleness channel, in which the captured half-spread rises by a factor of 1.33 on Bybit and 1.37 on Hyperliquid from the 100 ms rung to the 10,000 ms rung, and a fill-selection channel, in which the net moves as well, from -0.7323 to -0.4176 bp on Bybit and from -0.3435 to -0.0401 bp on Hyperliquid. Across all seven rungs on both venues the point estimate of the net is negative, though no rung carries an interval, and the Hyperliquid net at the coarsest rung is -0.0401 bp, smaller in magnitude than any two-way standard error reported for that venue. The sign is therefore stable across the tested range as a point estimate rather than established at each rung. How much of that range a reader should weigh is settled by the two cadences Section 8.1 measures, about 100 ms on Bybit and 541 ms on Hyperliquid. Across only the span those two cover, from the 100 ms rung to the 500 ms one, the net moves by 2.2% on Bybit and 0.5% on Hyperliquid. The 88% movement on Hyperliquid quoted above belongs to the full 100 ms to 10,000 ms sweep and not to that span, so the cadence the quoter actually faces is not where the rung sensitivity lives. The implication carried forward is that any statement about the share of a quantity coming from the capture leg is partly a statement about the publication cadence, and Section 8 attaches this qualifier where relevant.

$$\text{capture} = s(m_0 - f)/m_0 \quad (1)$$

$$\text{adverse} = s(m_\tau - m_0)/m_0 \quad (2)$$

$$\text{net entry markout} = \text{capture} + \text{adverse} = s(m_\tau - f)/m_0 \quad (3)$$

Across the seven re-quote gaps, the decomposition is reported in Table 2, weighted by fill count on the accompanying rung panel. The table gives capture, adverse, and net in basis points alongside fill counts for Bybit and Hyperliquid at gaps of 100, 250, 500, 1,000, 2,000, 5,000, and 10,000 ms, with the source file `reproduce/panels/venue_requote_rungs.csv` listed in the table note.

Bybit capture rises monotonically from 0.1732 bp at 100 ms to 0.2298 bp at 10,000 ms, while adverse moves from -0.9055 to -0.6473 bp and the net from -0.7323 to -0.4176 bp, with fills falling from 4,946,697 to 1,015,681. Hyperliquid capture rises from 0.2743 to 0.3757 bp, adverse from -0.6178 to -0.4158 bp, and the net from -0.3435 to -0.0401 bp, with fills falling from 1,484,203 to 408,599. The two venues are not directly comparable in this table because their calendars, coin sets, and tick sizes differ, a caveat that applies to every table in this paper. This re-quote sweep runs on 325 Bybit coin-days and 500 Hyperliquid coin-days drawn from 25 dates between 2025-07-27 and 2025-08-20, not on the headline panels, so its range is not a range the headline figures of Table 3 should be expected to fall inside; it measures cadence sensitivity within its own slice.

Proposition 1 states two invariances of the decomposition, and both are consequences of the definitions rather than empirical facts, so they are stated once as identities and then relied upon. The first is reference-mid cancellation: replacing the reference mid by any other positive value in both legs, holding the fill price and horizon mid fixed, leaves net unchanged up to a positive rescaling, so it cannot flip the sign. The second is horizon invariance of capture: capture contains no term evaluated after the fill, and is therefore

constant in the horizon, with every difference between a 10-second and a 60-second figure carried by the adverse leg alone.

Both invariances follow directly from (1)–(3), and neither requires any appeal to data. Their force is negative: (i) shows that any split reported anywhere in this paper is partly a property of the quoting cadence, whereas the sign of the net is not, and (ii) shows that any claim of invariance across horizons is not evidence about the captured half-spread. There is a scope restriction on (i): it holds for a fixed fill set, whereas changing the re-quote interval changes which fills occur at all, which is exactly why the net in Table 2 moves across rungs even though the reference mid cancels within every rung.

The split between capture and adverse depends on the choice of reference price even though the net is invariant to it for a fixed fill set, and this asymmetry is what the proposition formalizes. The proposition is stated as a numbered result because later sections depend on it directly: Section 5 reads the headline net as a measurement of this one maker, Section 8 uses the split only with the cadence qualifier attached, and Section 5 relies on the horizon invariance to attribute all 10-second-versus-60-second differences to the adverse leg.

That leaves the horizon invariance to be applied once more, and it is worth stating plainly. By part (ii) of the proposition, the captured half-spread is fixed at the moment of the fill, so it is identical in the 10-second and 60-second rows of every table that follows. Any claim that a result holds at both horizons is therefore a claim about the adverse leg itself, and Section 5 is phrased accordingly.

Table 2: The decomposition at seven re-quote gaps

gap (ms)	Bybit capture	Bybit adverse	Bybit net	Bybit fills	HL capture	HL adverse	HL net	HL fills
100	0.1732	−0.9055	−0.7323	4,946,697	0.2743	−0.6178	−0.3435	1,484,203
250	0.1853	−0.9122	−0.7270	3,741,866	0.2740	−0.6166	−0.3426	1,482,562
500	0.1921	−0.9084	−0.7163	3,178,371	0.2798	−0.6214	−0.3416	1,417,667
1,000	0.1988	−0.8999	−0.7012	2,594,186	0.3364	−0.6782	−0.3418	1,155,706
2,000	0.2065	−0.8860	−0.6796	2,047,175	0.3581	−0.6728	−0.3147	873,927
5,000	0.2183	−0.8545	−0.6362	1,415,084	0.3725	−0.6156	−0.2431	577,921
10,000	0.2298	−0.6473	−0.4176	1,015,681	0.3757	−0.4158	−0.0401	408,599

Trade-count-weighted requote-rung panel. For each gap from 100 to 10,000 ms, the table reports capture, adverse selection, net entry markout (all in basis points), and fill counts for Bybit and HL. Capture rises with gap on both venues, by a factor of 1.33 on Bybit and 1.37 on Hyperliquid; adverse selection and the net are negative throughout, with the net approaching zero at the longest gap. Source: reproduce/panels/venue_requote_rungs.csv

Proposition 1 (Two invariances). *Fix a set of fills, each with fill price f , maker side s , reference mid m_0 and horizon mid m_τ , and let capture, adverse and net be given by (1)–(3). Then: (i) Reference-mid cancellation. Replace m_0 by any other reference $m_0' > 0$ in both legs, holding f and m_τ fixed. The capture numerator moves by exactly $s(m_0' - m_0)$ and the adverse numerator by exactly $-s(m_0' - m_0)$, so their sum is unchanged and $\text{net} = s(m_\tau - f)/m_0'$, which differs from $s(m_\tau - f)/m_0$ only by the positive factor m_0/m_0' and therefore carries the same sign. A staler reference moves the boundary between the two legs and rescales their sum; it cannot flip it. (ii) Horizon invariance of capture. $\text{Capture} = s(m_0 - f)/m_0$ contains no term evaluated after the fill, so it is constant in τ . Every difference between a 10-second and a 60-second figure is carried by the adverse leg alone.*

5 The Central Result

The headline result is stated directly: on all three venues the captured half-spread is smaller in magnitude than the post-fill adverse drift. Consequently, the pooled 10-second pre-fee entry markout is negative. The two-way cluster interval by calendar month and coin excludes zero on every venue. Table 3 reports the decomposition, fill-weighted, in bp, with both intervals at 95%. The headline specification is the two-way cluster interval by calendar month and coin; the month-only interval is shown alongside so the difference is visible. The three panels cover different calendars, coin sets, and tick sizes, so the rows do not constitute a venue ranking.

Table 3: Pooled 10-second pre-fee entry markout

venue	capture	adverse	net	adv/cap	two-way CI	month-only CI
Bybit perpetuals	+0.2057	-1.0444	-0.8387	5.078	[-1.0835, -0.5939]	[-0.8692, -0.8083]
Binance USD-M	+0.4604	-0.8516	-0.3912	1.850	[-0.4516, -0.3308]	[-0.4213, -0.3610]
Hyperliquid	+0.5599	-1.0174	-0.4575	1.817	[-0.7286, -0.1864]	[-0.6733, -0.2418]

Fill-weighted, a mean over fills, in basis points; both intervals are 95%. The rows are not a venue ranking: the three panels cover different calendars and coin sets. Source: `reproduce/decomposition_by_venue.json`.

Every number in Table 3 is gross of maker fees, because the fee is a venue-specific constant that differs across panels; netting it out would mix fee-schedule differences into the microstructure measurement. Charging each venue’s published maker rate against the pooled pre-fee numbers gives -2.8387 , -2.3912 , and -1.9575 bp. The rebate that would pull each pre-fee number back to zero is 0.8387 , 0.3912 , and 0.4575 bp (by construction the negative of the net). Both of those calculations count costs only, and omit whatever the exit side of a round trip earns or costs, any rebate tier, and any other revenue. They therefore constitute no bound of any kind and are reported for completeness. The rebate dimension is run formally in Section 10.

The entry markout is one leg. A maker who completes a round trip posts again on the other side, and that second post has its own two legs, a captured half-spread and a drift against it. For the round trip to break even, the exit leg plus any rebate must together cover at least the entry shortfall: Bybit 0.8387 bp, Binance USD-M 0.3912 bp, Hyperliquid 0.4575 bp. This is an accounting identity about the measured quantity, not an estimate of the exit leg, which is not measured here. Nor should the identity be read as a reason to expect the exit to cover it. A passive exit at the touch is the same object this paper measures, pointed the other way, and on these panels that object is negative on all three venues; an exit that closes the gap would have to be something else, taken elsewhere in the book or aggressively, and neither is measured.

Adverse over capture runs from 1.8 to 5.1, and this section states plainly that it is the most quotable and least defensible number in the paper. Dividing by a captured half-spread that approaches zero on the highest-volume coins makes the ratio least stable precisely where the fill weight sits. Bybit BTC alone accounts for 34.5% of that venue’s fills, with a capture of 0.0124 bp and a coin-level ratio of 50.2. Thus, the pooled Bybit figure of 5.1 is not a mixture of the coin-level ratios but the ratio of the fill-weighted adverse sum to the fill-weighted capture sum. This is why it need not lie between the coin-level ratios at all and why the cross-coin medians differ from it. The cross-coin medians are 3.7, 1.7, and 1.1, none equal to the pooled value. Appendix A carries the per-coin columns. The

range across venues is reported because suppressing it would be worse, and it should not be read as an effect size. The absolute captures driving the highest coin-level values are small, at 0.0124 bp on Bybit BTC and 0.0237 bp on Bybit ETH, because the books on these largest names are one tick wide; a touch quoter earns roughly a half-tick per fill, and the measured captures sit inside the half-tick band those books imply: Bybit quotes BTC in 0.1 USDT and ETH in 0.01 USDT, which over the panel’s price range put a half-tick at 0.004 to 0.020 bp and 0.012 to 0.033 bp respectively. The ratio is therefore unstable where the fill weight is, and no shipped panel carries a quoted spread with which to check the point directly. This is a limit on the ratio rather than on the net: the denominator is small where the fill weight sits, so the ratio is unstable there, while the net it is built from is not. The books on these names are one tick wide for most of the window, so a capture near a half-tick is what a touch quoter earns rather than a sign of mismeasurement. Table 4 carries four cells per venue: the 10s and 60s horizons crossed with fill-weighting and coin-unit weighting. All twelve cells are negative, and ten of them exclude zero on the two-way interval. Limitation 1, two cells go untested: the two 60-second Hyperliquid cells have no two-way interval in the companion file, which records the fallback and returns no two-way verdict for them, so this table withholds them rather than reading the narrower month-only interval in their place. Limitation 2, sign invariant but magnitude not: on Hyperliquid the spread across the four cells is 1.06 times the magnitude of the smallest estimate itself (the choice of weighting and horizon jointly moves the answer by more than the answer), versus 0.15 on Bybit and 0.05 on Binance. Limitation 3, the coin dimension cannot move the verdict on two venues: on Bybit and Binance every coin agrees with the pooled sign, the closest to zero being -0.6094 bp (Bybit) and -0.0987 bp (Binance), so no coin resampling can reach zero; the cross-coin consistency is the evidence, not the interval. Hyperliquid is the only venue where the test could fail, and there the two 10-second cells pass and the two 60-second cells are withheld. On coin-unit weighting: it multiplies fill counts by average trade size; since posted size is fixed at one coin, one unit of one asset is not comparable to one unit of another. Within a coin it is a true size weighting; across coins it is not a value-weighted mean. Realised fill size varies between zero and one coin because the fill quantity is the overflow capped by the amount still resting, so the weighting reflects realised fills rather than the fixed submission size.

Table 4: The weighting and horizon family

venue	10s fill-wtd	10s coin-unit	60s fill-wtd	60s coin-unit	spread / smallest
Bybit perpetuals	-0.8387	-0.9614	-0.8343	-0.9345	0.15
Binance USD-M	-0.3912	-0.4011	-0.3838	-0.3879	0.05
Hyperliquid	-0.4575	-0.3075	-0.6344	-0.6025	1.06

Net entry markout in basis points, with four units per venue. Capture is fixed at the time of execution, so it is identical across horizons under the same weighting scheme; only the adverse-selection component varies with τ . All twelve units are negative. Ten of them exclude zero on the two-way interval; the remaining two, the 60-second Hyperliquid pair, have no two-way interval in the companion file, so their verdicts are withheld rather than failed. The rightmost column is the spread across the four units divided by the smallest magnitude among them. Source: `reproduce/decomposition_by_venue.json`

No single coin carries any venue. On Bybit, dropping BTC moves $-0.8387 \rightarrow -0.9595$ bp (14% of the effect). On Binance, dropping ETH moves $-0.3912 \rightarrow -0.3742$ bp (4%). On Hyperliquid, dropping ETH moves $-0.4575 \rightarrow -0.4035$ bp (12%). Every leave-one-coin-out value on every venue keeps its sign. Measured against the half-width of the month-only interval printed in Table 3, the Bybit swing is about four times

it, whereas on Binance and Hyperliquid it falls inside it, at roughly one half and one quarter respectively; another way to see that the month-only interval under-prices the coin dimension's contribution.

The adverse leg changes between 10s and 60s by a factor of 0.996 (Bybit), 0.991 (Binance), and 1.174 (Hyperliquid). Two reference values frame this. If the midprice were a martingale and order flow carried no information, the adverse term itself should be zero at every horizon; it is a signed markout, not an unsigned magnitude. If drift continued at its 10-second rate, the ratio should be six. The measured values sit far above zero and far below six, so the drift is substantially complete at ten seconds rather than still accumulating. The limitation is stated in the same breath: this compares one measured ratio against two reference values and does not test persistence; and since capture is fixed at the fill, the capture leg plays no part in the comparison.

5.1 Per Coin, Per Venue

Per-coin results are where the pooled headline numbers either hold together or fall apart, and the first fact to report is that they do not hold together in the same way across venues. On Bybit, all 13 coins are negative at both horizons and all 13 intervals exclude zero. On Binance, all 15 coins are negative at both horizons, 12 carry a clears verdict, XRP does not (net -0.0987 bp, interval $[-0.2022, +0.0047]$), and 2 verdicts are withheld. On Hyperliquid, 13 of 20 coins are negative at 10s and 17 of 20 at 60s; only 2 carry a clears verdict at 10 seconds (BTC and SOL), and 11 of 20 per-coin verdicts are withheld. The sentence “adverse selection over-consumes the touch on every coin” is therefore true at the two centralized venues and false, taken literally, at the on-chain one, because the pooled Hyperliquid number is fill-weighted. That venue's fill weight concentrates in BTC, HYPE, ETH and SOL, measured in simulated fills rather than venue volume, which no panel here reports. The 60-second per-coin surface is not tabulated here; its counts come from the shipped per-coin interval file `reproduce/per_coin_intervals.json`.

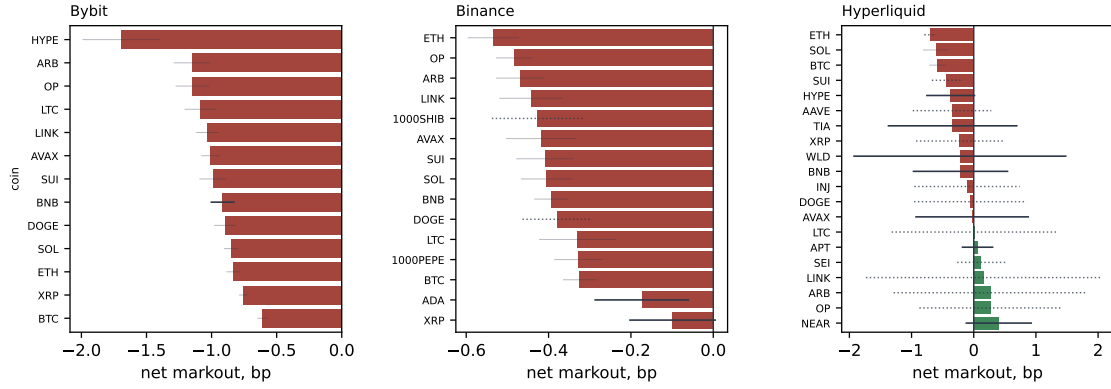


Figure 2: Per-coin net entry markout at the 10-second horizon, one panel per venue, most negative first. Whiskers are 95% month-clustered t intervals, the object the per-coin verdicts are read from, the pooled venue verdicts of Table 3 being read from two-way month-and-coin intervals; a faded whisker marks a cell whose month-resampling interval could not have reached zero, which is a property of that resampling rather than of the t interval drawn, and a dotted whisker marks a withheld verdict, of which there are none on Bybit, two on Binance and eleven on Hyperliquid. The three panels carry their own horizontal scales, spanning -2 to 0 bp, -0.6 to 0 and -2 to $+2$, so bar lengths are comparable within a panel and not across them. Hyperliquid is the only panel whose point estimates cross zero; on Binance, XRP’s interval crosses it while its point estimate does not. Source: `reproduce/decomposition_by_venue.json`, `reproduce/per_coin_intervals.json`.

Read as a description of dispersion rather than as 96 tests, the surface is what Figure 2 shows: that the per-coin intervals are overwhelmingly short relative to the spread of the point estimates across coins. Of the 96 per-coin cells (48 coins across two horizons), 57 carry month-resampling percentile intervals that could not have reached zero in the first place, because the underlying monthly means are of one sign throughout; there, consistency across months is the evidence and the interval adds nothing. Separately, 26 of the 96 cells have fewer than five effective months under the Kish effective-cluster count [22] applied to fill weights (below the level at which a clustered interval is worth reading), and these cells return no verdict at all, leaving 70 cells that do carry a verdict. This count of 70 overstates the evidence of exclusion: the two problems overlap, and together they mean that a reader who counts “cells that exclude zero” is running a family of tests whether the prose says so or not.

A skeptical reader will do the counting anyway, so we apply the multiplicity correction directly rather than argue about it. The primary family is the 10s cells only, since the two horizons for the same coin differ only in the adverse-selection leg and a 96-cell family would count each coin twice: of the 35 cells returning a verdict at 10s, 27 exclude zero at the raw 0.05 level, and all 27 survive Benjamini–Hochberg [10]. Carrying both horizons instead: 56 of 70 exclude zero and pass BH, of which 54 pass Benjamini–Yekutieli [11], which prices arbitrary rather than positive dependence. The two lost are both horizons of the same Binance coin, ADA, with raw values 0.011 and 0.012. Under either family, multiplicity is not what constrains this surface.

What does constrain it is the 26 cells that return no verdict and the 57 cells whose intervals could not have reached zero, p-value correction being powerless over both, so that we report the surface as dispersion. The correction results accompany `reproduce/per_coin_multiplicity.json`, with both families and both procedures, so the counting reading can be checked rather than trusted.

One cell needs both readings stated. Hyperliquid’s HYPE at 10 seconds has a t interval of $[-0.7538, +0.0118]$ that fails to exclude zero in both files, so its verdict is no, which is what Table 10 records. Undetermined is not the verdict but the percentile flag beside it. That interval’s upper bound sits within a hundredth of a basis point of zero against a point estimate of -0.3710 , so which side it falls on is settled in the fourth decimal and moves with the bootstrap stream: $+0.0022$ bp in the decomposition file and -0.0005 in the per-coin interval file, and a fresh rebuild of the first returns -0.0056 . The flag should therefore be read as undetermined rather than as a verdict of either kind, and the cell is the one documented exception the shipped-versus-rebuilt comparison carries. It is also why every verdict in this paper is read from the t interval rather than the percentile interval.

6 Trade-Sign Counterfactual

The preceding section’s per-coin verdicts rest entirely on the exchange-provided aggressor flags. Where an exchange does not publish such a flag, the direction must be inferred from prices and volumes, and the error thereby introduced cannot be priced without the very truth that is missing. Odders-White [27] and Andersen and Bondarenko [7] price that error against a perfect benchmark for stocks and for toxicity measures; what is extended here is the downstream quantity (a market maker’s entry markout, not a classification statistic), and it is done on venues where every fill carries its own flag.

Rerunning the same tape three times for each of the three exchanges, once with the exchange flag, once with Lee–Ready, once with the naive tick rule, yields the comparisons that follow. Lee–Ready classifies a fill as a buy when the trade price is above the prevailing midquote, a sell when below, and falls back to the tick rule when the trade is exactly at the midquote; here the prevailing midquote is the last quote strictly before the fill, not the original five-second lagged quote. The tick rule compares each price with the last different price and carries the previous direction through a run of equal prices. The entire simulation is rerun under each rule rather than re-labelling a fixed set of fills, because direction is not a label attached to a fixed set: it determines which side of the book a fill consumes, hence which of the market maker’s two resting orders could have filled, and only then does it enter the markout as the maker’s direction. A researcher with only a classifier obtains a different fill set and a different direction on each fill; Table 5 presents both channels.

The timestamp-collision share is 0.3% on Bybit, 8% to 10% on Hyperliquid, and 52.7% on millisecond-stamped Binance USD-M. On a tape, a fill and the book update it triggers may share a timestamp, in which case the last quote at or before the fill is not that quote; it is the post-fill book, and this inverts the classification precisely on the price-moving fills the rule is designed to catch. This is a property of the tape’s clock, not of the classifier. Read at or before the fill, the quote rule scores 0.620 on one heavy Binance day; read strictly before the fill, it scores 0.996; Bybit gives 0.926 under either reading. Every number in this section takes the last quote strictly before the fill. A study adopting the other reading is reporting its venue’s clock granularity as a property of Lee–Ready, and would find the three venues incomparable for a reason that has nothing to do with the classifier. The timestamp-collision shares and the two quote-rule scores in this paragraph are the only numbers in the measurement sections read directly from the raw tape rather than from the shipped package; everything else is either in a results file or computed from a shipped coin-day panel, because the raw archive lies outside the boundary of Section 13.

Section 14’s listing counts of 206 and 185, and the death rates derived from them, are the other place the paper steps outside the package.

Table 5: The same tape under the exchange flag and two classifiers, each taking its own fill set

Venue	Rule	Agreement rate	Capture	Adverse selection	Net	Net error	95% CI on the error	Fills, $\times$ truth
Bybit	exchange flag	1.000	+0.2057	−1.0444	−0.8387	—	—	1.000
Bybit	Lee–Ready	0.932	+0.2163	−1.0266	−0.8103	+0.0321	[−0.0001, +0.0643]	1.060
Bybit	tick rule	0.898	+0.2006	−0.9966	−0.7960	+0.0318	[−0.0018, +0.0654]	0.937
Binance	exchange flag	1.000	+0.4604	−0.8516	−0.3912	—	—	1.000
Binance	Lee–Ready	0.982	+0.4606	−0.8488	−0.3882	+0.0025	[+0.0015, +0.0036]	0.998
Binance	tick rule	0.913	+0.4589	−0.8018	−0.3428	+0.0523	[+0.0169, +0.0876]	1.011
Hyperliquid	exchange flag	1.000	+0.5599	−1.0174	−0.4575	—	—	1.000
Hyperliquid	Lee–Ready	0.936	+0.5183	−1.0834	−0.5651	−0.0931	[−0.1243, −0.0620]	1.082
Hyperliquid	tick rule	0.894	+0.5467	−1.0841	−0.5374	−0.0527	[−0.1162, +0.0107]	0.930

10-second horizon, trade-count weighted, in bp. Agreement rate = share of trades, weighted by count, whose direction matches the exchange. Net error = the paired within-coin-day difference between the rule’s net and the exchange flag’s net, which is the estimand the interval beside it is built on and is therefore not the difference of the two Net cells, those being pooled, so a positive error means the rule reports a value less negative than the truth. The interval on this error is paired: differences are taken within each coin-day, then clustered, taking the wider of the month-only and month-plus-coin two-way intervals. Fills = the rule’s fill count as a multiple of the true value. Source: `reproduce/sign_rule_counterfactual.json`

Table 5 delivers six readings from the agreement rates and error magnitudes. First, the classifiers are wrong about two to eleven trades in a hundred. Lee–Ready agrees with the exchange on 93.2% of Bybit fills, 98.2 on Binance, 93.6 on Hyperliquid; the tick rule gives 89.8%, 91.3% and 89.4%. Weighting by trade size rather than count raises all six figures, by between 0.2 and 3.1 percentage points, using the size-weighted sign accuracy recorded per coin-day in the sign-rule panel, so the misclassified fills are concentrated among the small ones. The cross-coin dispersion is far wider than the pooled numbers suggest. The venues differ in how wide: tick-rule agreement runs from Bybit OP’s 86.1% to Bybit BTC’s 94.6%, from Binance ARB’s 84.8% to Binance BTC’s 97.3%, and from Hyperliquid TIA’s 74.2% to Hyperliquid BTC’s 95.0%. The pooled agreement figures are not a good guide to any particular coin.

Second, every rule keeps the sign of the result on every venue. Under all three rules on all three exchanges, the captured half-spread is smaller in magnitude than the adverse-selection drift and the pooled net markout stays negative, so a study with only a classifier would reach the same conclusion. This is the one thing readers of the equity literature most need to know about their own markout results on venues whose feeds carry no aggressor field and cannot verify on their own data.

Third, the magnitude moves, and its direction is not shared. On Bybit and Binance both classifiers report a net markout that is less negative than the truth (+0.0321 and +0.0318 bp on Bybit, +0.0025 and +0.0523 bp on Binance), which is 3.8% and 3.8% of Bybit’s measured effect and 0.6% and 13.4% of Binance’s. On Hyperliquid both report a more negative markout, by −0.0931 and −0.0527 bp, or 20.4% and 11.5%. Two venues sit on one side of zero and one on the other, so no constant correction exists, and a study that assumes one is applying a centralized-venue adjustment to an on-chain panel. The portable statement is a bound, not a correction: across these three panels, a classifier moves the pooled entry markout by no more than a quarter of its own magnitude and does not change its sign.

Fourth, accuracy does not predict the bias across venues. Within a venue, the more accurate rule delivers the smaller error on only one of the three, and the two Bybit rules tie.

Across venues the ordering fails outright. Hyperliquid’s Lee–Ready is more accurate than Bybit’s (93.6 vs 93.2), yet its error is about five times as large a share of its own effect (−20.4 vs +3.8%). Binance’s tick rule is more accurate than Bybit’s (91.3 vs 89.8), yet its share is more than three times larger (+13.4 vs +3.8%). Binance under Lee–Ready is the most accurate cell in the table and also the smallest error, but that pairing does not transfer to another venue. Classifier agreement is therefore not a sufficient statistic for a classifier’s effect on markout, and reporting one without the other overstates what is known.

Fifth, half the errors clear zero. Three of the six paired intervals exclude zero; three contain it. Both Binance cells exclude zero under 11 month clusters and 8.4 effective clusters; Hyperliquid under Lee–Ready excludes zero under 11 month clusters and 6.3 effective clusters; both Bybit cells contain zero, and they contain it only just, with lower bounds of −0.0001 and −0.0018 bp against upper bounds of +0.0643 and +0.0654 bp. These are point estimates that are order-stable but unevenly significant: the section reports how large the error can be, not that it is reliably nonzero everywhere.

Sixth, part of the error is a different fill set, not a different sign. A rule that misassigns direction can create or destroy a fill, not merely mislabel it. Across the six cells, the fill count relative to the truth moves between 0.930 and 1.082, and the capture and adverse-selection terms in Table 5 both move; this can happen through two channels, a changed fill set and a flipped sign on a retained fill, since capture is the signed difference between the midprice and the fill price divided by the midprice. Any correction coefficient read from this section corrects both channels at once.

The agreement rates here are agreement with the exchange flag, which the paper treats as the truth; where the exchange’s own flag is wrong, this measures agreement, not correctness. The numbers are for a single quoter and a single re-quoting cadence; the cost of a classifier under a different quoting rule is not measured. Three venues are three points: they establish that the error is not constant and that agreement does not rank it, but what the error depends on is not identified here.

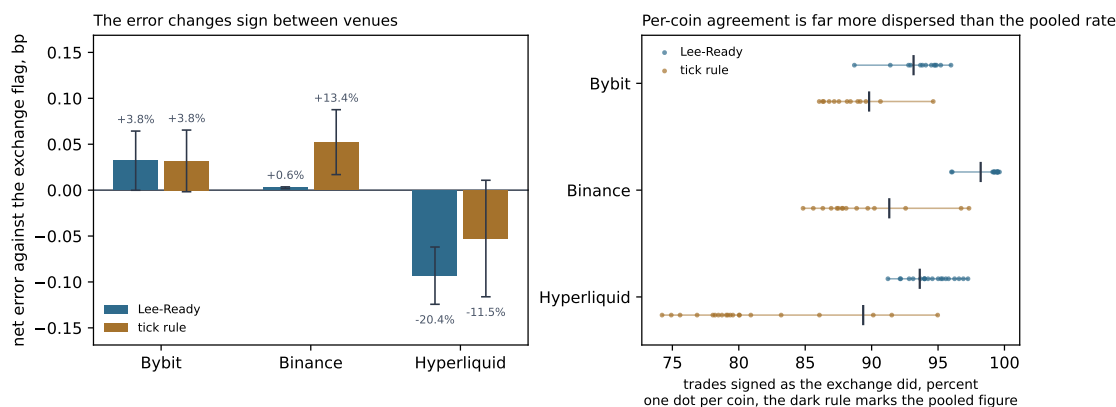


Figure 3: What a trade-sign classifier costs, and why no constant correction exists. Left: the paired net error each rule makes against the exchange flag, by venue, with the two-way intervals of Table 5; each bar is labelled with that error as a share of the venue’s own measured effect, +3.8% and +3.8% on Bybit, +0.6% and +13.4% on Binance, and –20.4% and –11.5% on Hyperliquid, so the two centralized venues sit on one side of zero and the on-chain one on the other. Right: per-coin agreement with the exchange flag, one point per coin for each rule, with the pooled value marked; tick-rule agreement spans 8.6 points on Bybit, 12.5 on Binance and 20.7 on Hyperliquid, and 9.9, 14.8 and 23.0 across both rules together. Read together, Hyperliquid’s pooled Lee–Ready agreement is the higher of the two while its error is about five times the larger share, which is why agreement does not rank the error. Source: `reproduce/sign_rule_counterfactual.json`

Figure 3 pairs the two channels. The left panel plots each rule’s paired within-coin-day net error against the exchange flag, with the two-way intervals of Table 5 and the error as a share of that venue’s own measured effect; the two centralized venues sit above zero and the on-chain one below, so the section reports a bound rather than a correction. The right panel plots the same two rules’ per-coin agreement with the exchange flag, one point per coin, with dark tick marks for the pooled values; the dispersion is far wider than the pooled numbers suggest. Reading the two panels together is what shows that agreement does not rank the error: Hyperliquid’s Lee–Ready point lies above Bybit’s, while its bar sits about five times farther from zero.

7 Inference

The verdicts above rest on a single interval choice; this section reports what that choice is, what it costs, and where the entire family of alternatives breaks down together.

7.1 Months alone

Because no single natural unit exists for a coin-day panel, the choice of clustering unit is not innocuous. Clustering by calendar month prices common time-series shocks; clustering by coin prices the fact that days of the same coin resemble one another. The two are not nested, so we adopt the two-way estimator of Cameron, Gelbach and Miller [13], which prices both dimensions simultaneously.

Table 6 reports standard errors on the pooled 10-second net under months-only and two-way clustering. Crossing both dimensions widens the standard error at every venue, and no verdict changes. The Bybit ratio of 6.0 is the most striking: with 29 month clusters

and 13 coins, trade-weight concentration is sufficient that the coin dimension carries most of the residual dependence, and the months-only interval prices none of it. Reporting the months-only interval as the headline for Bybit would make that estimate look six times more precise than the panel supports. Every verdict survives this correction, which was applied before verdicts were read, not after. For every headline interval, the t multiplier is taken at one fewer than the smaller of the two effective cluster counts, computed by the Kish formula. The effective month counts are 23.2, 8.4 and 6.3 on Bybit, Binance and Hyperliquid; the effective coin counts, recovered from the per-coin fill shares of Table 10, are 5.96, 10.73 and 5.43. The binding dimension is therefore the coin count on Bybit and Hyperliquid and the month count on Binance, giving multipliers of 2.78, 2.36 and 2.78. Two of the three headline intervals thus rest on about five effective clusters, just above the floor below which this paper withholds a verdict, and one notch above the 4.4 effective coins at which the venue-difference verdict of Section 8 is withheld. The shipped file records the effective month counts but not the effective coin counts, so those were recovered from the shipped per-coin fill counts rather than read from the package.

Two further limits attach to these intervals. First, the accompanying quantile intervals resample calendar months and under-cover at 29, 11 and 11 clusters, so verdicts are read from the t -intervals beside them. Second, the two Hyperliquid 60-second cells in the accompanying file cannot be constructed two-way. They fall back to the months-only standard error, and the file marks them: `net_two_way_interval_was_constructed` is false and the two-way verdict is null rather than a pass read off the narrower interval, which is why Table 4 withholds them.

Table 6: Standard error on the pooled 10-second net

Venue	Month-only SE	Two-way SE	Ratio	Two-way verdict
Bybit perpetuals	0.0147	0.0882	6.0	clears zero
Binance USD-M	0.0128	0.0255	2.0	clears zero
Hyperliquid	0.0839	0.0977	1.2	clears zero

The table compares month-only clustering with month-and-coin two-way clustering by venue. Standard errors are reported for each specification, along with their ratio and the two-way verdict. Crossing both dimensions widens standard errors at every exchange without changing any verdict. Source: `reproduce/decomposition_by_venue.json`.

7.2 Interval calibration

The calibration exercise runs the same pipeline across four estimators under twenty-one generative conditions, then repeats the two conditions that decide the interval choice in this paper’s own clustering configuration, and adds a month sign-flip test calibrated separately on each exchange’s own month structure. The sources are `reproduce/coverage_sweep.json`, whose `paper_configuration` block carries the second study, and `reproduce/wild_test_calibration.json`. Three results enter the verdicts.

First, two-way clustering earns its width, but not for the reason the main grid suggests. That grid clusters one-way on the coin and crosses in the date, whereas this paper clusters one-way on the calendar month and crosses in the coin, so the grid cannot speak to the choice made here: a coin effect is the dimension its own one-way estimator already absorbs, and both estimators sit at 0.060 there. Measured in this paper’s configuration instead, on a panel shaped like Bybit’s at 29 months and 13 coins, a shock common to every coin on one date leaves both estimators at nominal, 0.052 either way, because dates

nest inside months and month clustering already absorbs it. What month-only clustering cannot see is a coin effect that persists across months: there it rejects a true null on 0.867 of replications against a nominal 0.05, while the two-way estimator holds at 0.057. On a panel shaped like Hyperliquid’s at 11 months and 20 coins the same pair reads 0.040 either way and 0.777 against 0.048. Table 6 therefore rests its verdicts on the wider number, and the reason is the coin dimension rather than the calendar one. Second, on the question the main grid does answer, the condition that breaks every method at once is weight concentration, not skew. When the heaviest clusters are also the atypical ones, which is what Bybit BTC looks like at 34.5% of that venue’s fills on a capture of 0.0124 bp, the t and two-way estimators run at a size of 0.183 against a nominal 0.05, and the guard abstains on 86% of replications, so only 14% return a verdict at all. Two-way clustering does not repair this, because the problem is which observations carry the weight rather than an unpriced second dimension. Skewed coin effects are milder than the grid’s labels suggest. The grid varies the sigma of a lognormal, not a skewness: sigma 0.44 is a skewness of 1.5, at which all four methods sit near nominal at 0.045 to 0.093, and sigma 0.55 is a skewness of 2.0, at which they sit at 0.057 to 0.100. No coin-effect skewness is measured on these panels, so the comparison is made against the most right-skewed distribution that is measured, Hyperliquid’s across calendar months at +2.02, with Bybit at −0.68 and Binance at +0.45. That worst case lands on the sigma 0.55 row and is still near nominal. The size of 0.42 that the grid reports belongs to sigma 2.0, whose implied skewness is 414, two orders of magnitude beyond anything measured here. Third, the sign-flip test lands near nominal size on two of the three panels and not on the third. Calibrated on each exchange’s own months: 0.053 Bybit, 0.061 Binance, 0.124 Hyperliquid (the last driven by a cross-month distribution skewed right at +2.02, with 73% of its months falling on the same side of the median). The latter two share a cluster count of 11, so cluster count is not what separates them. No verdict in this paper relies on that test; the calibration is reported so that readers who might adopt it know what that adoption costs.

7.3 Withheld verdicts

Where the Kish effective cluster count falls below the level at which a clustered interval is worth reading, on whichever of the clustering dimensions is the smaller, the accompanying files return a verdict of withheld rather than a weak one. On the per-coin cells that dimension is the month, because they carry no second one; on the pooled cells and on the survivorship contrast it can be the coin. That guard withholds 2 of 15 Binance coins and 11 of 20 Hyperliquid coins, for 26 of 96 per-coin cells across both horizons, and it also makes the per-coin clustered verdict for the venue difference in Section 8 unavailable, at 4.4 effective coins of 13. Withheld is not a failed test, and the counts distinguish the two wherever they appear: the 96 per-coin cells split 26 withheld against 14 that ran and did not clear. What the guard catches and what it misses are both measured: it triggers on concentration, because it is weight-based, and it is blind to the coin-effect skew of Section 7.2; `reproduce/coverage_sweep.json` records how often it returns a verdict in its `verdict_available_frac` column.

8 Venue Comparison

Two same-date venue comparisons are available in these panels. This section reports the Bybit against Hyperliquid one, because it is the pair the shipped interval machinery covers, and runs it on the 325 coin-days where both exchanges traded on the same date, balanced as 13 coins across 25 dates and taken from the 2,000 ms re-quote panel. The pooled rows of Table 3 cannot be differenced, because the three panels cover different calendars and coin sets; this same-date design is one of two subtractions the data admit. The other, Bybit against Binance, shares 1,284 same-date coin-days across 12 coins and 107 dates, giving a pooled difference of -0.4883 bp and a paired difference of -0.5181 bp, with Bybit the worse of the two on 1,228 of 1,284 coin-days and on all 12 shared coins; we report that overlap here as a descriptive check that carries no interval and no verdict. No shipped analysis covers that pair, so those five figures are computed directly from the two shipped coin-day panels rather than read from a result file. The same-date estimate is -0.3546 bp pooled and -0.3455 bp paired, under the convention that negative means Bybit's net is the lower of the two. The sign is negative for 13 of 13 coins, with the closest to zero at -0.2317 bp, and Bybit is the worse venue on 293 of 325 coin-days.

Direction is the claim and magnitude is not, for four reasons that are each larger than the interval. First, the estimate moves with the sampling scheme it is drawn from. The comparison uses the re-quote panel; the headline numbers of Table 3 use the two main panels. The re-quote panel covers the same 25 consecutive dates on both venues, 2025-07-27 to 2025-08-20, but the two main panels reach into that window to different depths, so the coin-days on which the two schemes can be compared at all differ by venue. All 25 dates sit inside Hyperliquid's main panel, whose consecutive block of Section 3 covers that stretch and one day beyond it, giving 500 shared coin-days on 20 coins. Only 12 of the 25 sit inside Bybit's, whose three-day stride ends on 2025-08-18, giving 156 shared coin-days on 13 coins. Agreement on those shared coin-days is what differs. Bybit's 156 coincides numerically with the 156 coin-days its main panel shares with Hyperliquid's, both being 13 coins on 12 dates, and the two are different intersections. The Bybit values agree on all 156, the two extractions returning the same net to nine decimal places of a basis point. The Hyperliquid values agree on 3 of 500, and the disagreement is not confined to the capture leg. On those 500 coin-days the main panel carries 729,383 fills against the re-quote panel's 1,484,203, a capture of 0.4514 bp against 0.2743, and a net of -0.5825 bp against -0.3435 . That is a gap of 0.2390 bp on a venue whose headline figure is -0.4575 bp, and the main panel's capture sits above even the 10,000 ms rung's 0.3757, so the two extractions were cut differently by more than snapshot spacing alone. Because they carry different fill sets rather than the same fills priced against a different reference mid, Proposition 1 does not apply and the net is not protected by it. That divergence is unexplained. It is carried as a threat in Section 14 rather than resolved, it falls on the venue this paper already calls the weakest of the three, and the two extractions are reported leaf by leaf in the shipped file so a replication can start from the discrepancy rather than rediscover it. On coin-days covered by both schemes, the gap reads -0.3496 bp one way and -0.2232 bp the other: the sign holds while the magnitude moves by more than a third. The two rates therefore have different bases and are not read against a common denominator.

Second, the rung is hand-fixed, not derived. The 2,000 ms rung is hand-fixed. The accompanying re-quote intensity statistic cannot separate its own top two candidates, so it selects no rung at all. What it does say is that the reported rung is not among them.

Reported across all seven rungs in Table 7, the net difference is negative at every rung, ranging -0.383 to -0.352 bp, while the share of the difference coming from the capture term moves from 0.39 to 0.62. This rung sweep uses the same matched shared-coin-day estimator as the headline venue difference. Its capture share is therefore not the difference of the two venue rows in Table 2. Those rows are computed on each venue’s own unmatched rung panel. The direction is rung-invariant; its decomposition into the two terms is not.

Table 7: Bybit-minus-Hyperliquid same-date difference across re-quote rungs

Gap (ms)	100	250	500	1,000	2,000	5,000	10,000
Pooled difference (bp)	-0.383	-0.379	-0.368	-0.352	-0.355	-0.382	-0.362
Paired difference (bp)	-0.376	-0.363	-0.347	-0.335	-0.346	-0.369	-0.349
Share from capture	0.394	0.390	0.420	0.592	0.611	0.584	0.617

Negative values indicate that Bybit’s net is the lower of the two. The net difference keeps its sign at every rung, while the share attributable to capture rises from 0.39 to 0.62 across rungs. Source: reproduce/venue_difference.json

Third, the dispersion across estimators exceeds the date-clustered interval on them. Six settings of the same comparison, differing only in how coin-days are weighted, return five distinct values across those six settings (two of the six settings coincide). These five distinct values span -0.3455 to -0.4960 bp, a spread of 0.1505 bp, which is 2.6 times the width of the date-clustered interval. All six settings agree in sign. The reported pooled value of -0.3546 bp sits inside that span of the five distinct values rather than at either end, and the smallest-magnitude member of the family of the six settings is the paired estimate of -0.3455 bp.

Fourth, neither clustering dimension supports a verdict. The 25 date clusters are consecutive days within a 3.5-week window, not independent draws; the lag-1 autocorrelation is 0.64, leaving roughly 5.5 effective clusters, so the date interval is about 2.1 times too narrow. The exact binomial test sits at its own floor, saying at most that all 25 days share a sign. On the coin side, trade-weight concentration makes 13 coins behave like 4.4 by the Kish count, below the threshold at which an interval is worth reading, so that verdict is withheld, not reported as passed. The two-way interval by coin and date is $[-0.4300, -0.2611]$ at three degrees of freedom, and its verdict is likewise withheld. The evidence for the direction is the cross-coin consistency and the 293-of-325 coin-day count, not any of these intervals.

What the difference is made of, and what it is not. On the headline estimator, about 61% of the gap comes from the capture difference, not the adverse-selection difference; of the five weightings the composition is computed under, four attribute the majority to capture, with shares ranging 0.34 to 0.70. The reading is that Hyperliquid earns a wider effective spread (a tick-size and depth property) rather than facing less toxic flow. That reading itself depends on the rung (Table 7), so it is offered as the majority reading across weightings, not a measured decomposition. The pooled estimate of -0.3546 is the value this section opens with and the results inventory carries, whereas the paired estimate of -0.3455 is the smallest-magnitude claim in this paragraph; the two are distinct and should not be conflated.

No causal claim is made. The two exchanges differ simultaneously in matching architecture, tick grid, fee schedule, and publication frequency, and nothing here separates them. It is not established, and not claimed, that exchange architecture caused the observed difference. The comparison is also gross, before maker fees and without funding rates.

Bybit charges 2.0 bp and Hyperliquid 1.5 bp. The 0.5 bp fee-schedule gap is therefore larger than the 0.35 bp microstructure gap measured here, and it points in the same direction.

8.1 Lead-lag

The lead-lag analysis runs on the same 325 shared coin-days, cross-correlating returns on a 250 ms grid. The median peak places Bybit ahead by 500 ms, with Bybit leading on 322 of 325 coin-days and 13 of 13 coins. The date-clustered interval at this median collapses to zero width on this grid, so no verdict is reported rather than a false one.

This result cannot be read as informational leadership. The two exchanges publish at different frequencies, 100 ms against 541 ms, a 441 ms gap of the same magnitude as the 500 ms lead, and this alone could make the slower feed appear to follow the faster one. The panel carries only 12 ms of cadence variation across Hyperliquid coin-days, insufficient to test this confound. Within that narrow variation, the correlation between cadence and the measured lead is -0.485 , a sign opposite to the confound's prediction (a slower feed lagging for mechanical reasons overstates the lead, not understates it), which is suggestive but cannot exclude the confound. The honest statement is that the panel cannot separate informational leadership from publication-frequency differences, and no informational-leadership claim is made. Source: `reproduce/cross_venue_leadlag.json`.

9 Conditional Decomposition

We test whether the decomposition is invariant across volatility and session states by fixing those states and comparing the high-minus-low contrast of net entry markup. We split a Bybit subsample into eight-hour UTC blocks (Asia 00–07, Europe 08–15, US 16–23) and strictly causal volatility states, then run a cross-coin sign test on the high-minus-low contrast of net entry markup. The subsample is its own draw and not the headline panel: 1,176 coin-days on 12 coins across 98 dates, carrying 5,982,695 fills against the headline panel's 3,618 coin-days and 45,476,018 fills, and 94 of its 98 dates fall inside the headline panel while 4 do not. Its levels are therefore not comparable with Table 3, and only the within-block contrast is read from it. State labels come from terciles of realized volatility over the 30 minutes preceding each fill, computed from midprices strictly earlier than that fill, with each date's cutoffs taken only from fills on strictly earlier dates. Dates with insufficient history are marked unavailable and dropped rather than defaulted to the middle tercile, which removes 3.3% of the fill flow; that share is measured upstream, since the dropped cells do not survive into the shipped conditional panel and cannot be recounted from it.

The thirteenth Bybit coin is absent because HYPE listed midway through the window and carries only 90 coin-days in the headline panel against 284 to 295 for every other coin, too little history to support a tercile split. The contrast is positive for 7 of 12 coins in Asia, 7 of 12 in Europe, and 9 of 12 in the US block. Blocked sign-flip tests with block length five give $p = 0.789, 0.772, 0.285$, respectively, and none of the block lengths tried, 1, 3, 5 and 10, passes 5%. Figure 4 plots the share of coins with a positive contrast by UTC block, with counts and the block-five p values annotated, and its caption states the point directly: this is a failure to reject on 12 coins, not a demonstration that the result is

unchanged across regimes.

Under the sign test on 12 coins, a moderate shift cannot be detected, so no block shows a detectable calm-to-stress change in net entry markout under this test. This is not a statement that netting is invariant across volatility states, nor evidence of invariance, and the wording throughout is chosen accordingly; a failure to reject does not constitute support for the null.

First, the null distribution imposes a shared sign on every date within a block, which by construction carries any same-day dependence across coins into the null without estimating or correcting for it. An alternative path through a 66-pair correlation matrix left two-thirds of the pairs with insufficient overlap, and zero-filling those pairs, which asserts independence, produced spuriously significant results; we record that path as rejected. Second, only the net contrast is reported, because the gross capture leg involves choices about weighting and which cells to include, and those choices are not free, so we report no p-value for that leg.

This analysis speaks only to this Bybit subsample; it says nothing about the rest of that venue's panel or about the other two venues.

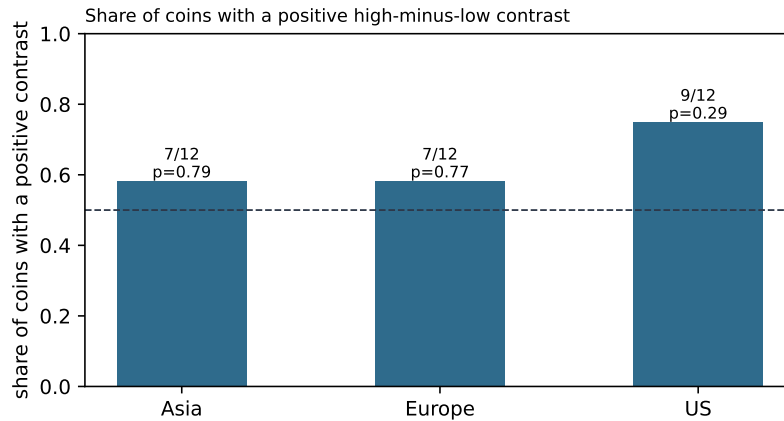


Figure 4: Share of coins whose high-minus-low net entry markout contrast is positive. Three bars show the share of coins with a positive contrast for Asia, Europe, and US, with counts and block-length-five sign-flip p-values: 7/12 ($p=0.79$), 7/12 ($p=0.77$), 9/12 ($p=0.29$). This is a failure to reject across 12 coins, not evidence that results are invariant across regimes. Source: `reproduce/conditional_net_asia.json`, `reproduce/conditional_net_europe.json`, `reproduce/conditional_net_us.json`

10 Depth and Rebate Frontier

Whether quoting deeper helps is the natural question. We answer it on a 300 coin-day Bybit buy-side subsample, measured at five price levels from the mid, at the ten-second horizon, averaging over filled rows. The answer is that each fill improves while fills become rarer. Per quoting opportunity the two effects push the same way for a losing quote, since a lower fill rate on a negative markout loses less, which is why quoting deeper is the least bad of the five at zero rebate and why the per-opportunity ordering has to be read against a rebate paid per fill. Measured per quoting opportunity, the deepest rung is roughly eighteen times better than the touch. The ordering therefore turns on the size of the rebate paid per fill rather than on the two effects cancelling. This per-opportunity

ordering rewards not trading, so the comparison is between rungs and not against staying out.

Table 8: Depth frontier

Price rung from mid-price	1 (touch)	2	3	4	5 (deepest)
Net markout per fill (bp)	−0.7171	−0.6535	−0.5128	−0.4513	−0.3283
Fill rate	0.003348	0.001075	0.000700	0.000518	0.000408
Number of fills	610,366	195,945	127,570	94,369	74,462
bp per quote opportunity	−0.002401	−0.000702	−0.000359	−0.000234	−0.000134
Break-even rebate (bp)	0.7171	0.6535	0.5128	0.4513	0.3283

For 300 Bybit coin-days, buy side, 10-second horizon, averaged over executed rows. Tiers are price tiers from the mid-price, not queue positions. Before any rebate, each tier is negative; the break-even rebate declines with depth, while the fill rate declines faster. Source: `reproduce/depth_rebate_frontier.json`.

Table 8 reports the frontier. Rungs are price rungs from the mid, not queue positions. Every rung is negative before any rebate; the break-even rebate falls with depth, whereas the fill rate falls faster. Per fill, the touch sits at −0.7171 bp and the deepest rung at −0.3283 bp, a gap of 0.3888 bp in favor of the deepest rung. For each quoting opportunity (the denominator a maker actually faces when choosing where to post), that ordering holds only while the rebate does not exceed 0.7711 bp. Above that threshold, the touch’s far higher fill rate amplifies the rebate faster than its worse markout subtracts, so the touch wins.

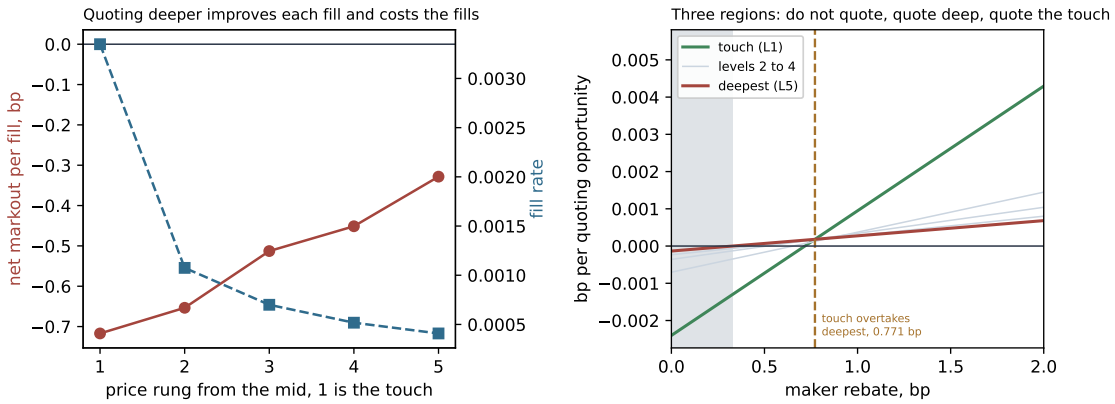


Figure 5: Two panels. Left: net markout per fill improves monotonically with depth across the five price levels measured from the mid, while the fill rate falls to 12% of the touch at the fifth level, so quoting deeper improves each fill and costs fills. Right: expected value per quoting opportunity against the maker rebate, with a shaded band marking where no level is profitable and a dashed line at 0.771 bp where the touch overtakes the deepest level. Three regions: do not quote, quote deep, quote the touch. No rebate tier is drawn: the shipped interval file withdraws its published-tier attribution as unsourceable. Source: `reproduce/depth_rebate_frontier.json`

Figure 5 shows the frontier. The left panel plots per-fill net markout, which improves monotonically with depth, while the fill rate at the fifth level falls to 12% of the touch’s. The right panel plots per-quoting-opportunity expected value against the maker rebate, with a shaded band marking the region where no level is profitable and a dashed line at the rebate where the touch overtakes the deepest level (0.771 bp). Three regions partition the frontier: below 0.328 bp no level is profitable per quoting opportunity; between 0.328

and 0.771 bp the deepest level is the best of the five; above 0.771 bp the touch overtakes it, because its far higher fill rate amplifies the rebate faster than its worse markout subtracts.

Interval evidence on the crossing comes from `reproduce/depth_rebate_intervals.json`, not from the frontier file that carries the point estimate. The crossing is 0.7711 bp, with a coin-clustered interval [0.7163, 0.8341] and a coin-day interval [0.6753, 0.8698]. It is finite on all 13 coins, with median 0.812 and range 0.591 to 1.102; leave-one-coin-out moves it only within [0.7597, 0.7959].

A residual fee on top of the rebate shifts the crossing one-for-one, because the maker fee is charged in addition to the rebate. A residual maker fee of 0.5 bp moves the crossing from 0.7711 to 1.271 bp, and a fee of 1.0 bp moves it to 1.771 bp. The crossing is a property of this panel, whereas whether a given maker's rebate sits above or below it is a property of that maker's fee schedule, a dimension the bootstrap does not cover.

No tier is asserted here. Section 12 restates the same crossing and observes that which side a maker falls on is decided by that maker's schedule. That schedule carries monthly uptime, product coverage, and minimum order size obligations, and a single-rung, single-instrument, buy-side-only simulated quoter is not shown to satisfy them. It is an upper bound on what such a quoter could earn.

Three further limits apply. The subsample is 300 of 3,618 panel coin-days, or 8.3%, and it is unevenly distributed (45% of it falls in 2023, against 28.9% of the remainder). Separately, coin-level structure is inconsistent: only 2 of 13 coins are monotonic across all five levels, 8 of 13 are negative at every level, and 5 of 13 turn non-negative at some depth rung; these five are cells a replication should look at first, since 11 of 13 have a deepest level better than the touch, while on LTC and SUI deeper is worse. The crossing itself moves from 1.1301 to 0.7711 bp after one fill-model correction, and only the corrected value is reported.

11 The Quoting Arm

The depth rule tested in Section 10 is not the only quoting strategy a maker could run, and the natural alternative is one that accounts for inventory. On 91 Bybit coin-days across 13 coins, we test an inventory-aware depth rule against the touch, scored at the 10-second horizon with a planned 300-second unwind. The 300-second unwind is part of the quoting rule being tested, not a priced exit: the arm plans its inventory over that horizon, and no exit cost is charged. It is then scored on the same 10-second entry markout as everything else, so the unwind does not enter the reported figures as a cost. The -0.5439 bp figure is therefore an entry markout, comparable with the touch on that basis. Per fill, the tested arm is better than the touch: -0.5439 versus -0.9098 bp, a difference of $+0.3659$ bp. Table 9 reports the comparison, showing that it does not escape adverse selection: it is still negative on 66 of 91 coin-days, and it achieves its improvement by taking only 9.2% of the fills the touch rule takes. In this comparison, the touch figure is re-weighted with the quoting arm's weights; the arm-weighted touch net of -0.9098 bp is therefore not the same object as the headline -0.8387 bp or the depth frontier's -0.7171 bp, which are computed on different samples and weights.

Table 9: The tested quoting arm against the touch

	touch rule	tested arm
Net entry markout per fill (bp)	−0.9098	−0.5439
Fills relative to the touch rule	1.000	0.092
Coin-days with negative net entry markout	90 of 91	66 of 91
Break-even rebate (bp)	0.8742	0.5439
Better per quoting opportunity at 0 bp rebate	no	yes
Better per quoting opportunity at 1.5 bp rebate	yes	no

91 Bybit coin-days. The quoting arm is better per fill but fills far less often, so the decision-relevant comparison is the rebate frontier, since a rebate is paid per fill; the last two rows are per quoting opportunity and the crossing between them sits at 0.9076 bp. Footnote: the touch-price row's per-trade net markout and its break-even rebate are not mirror images of the quoting arm's, because the former is weighted with the quoting arm's weights, while the latter is the negated touch-price net markout weighted by touch-price trade counts; both describe the same quantity, differing by 0.036 bp. Source: reproduce/avellanedastojkov_arm.json

Three qualifications decide how much this is worth. First, significance depends on the clustering choice. Clustering by coin-day gives $[+0.1754, +0.5563]$, an interval that excludes zero; clustering by coin gives $[-0.0109, +0.7427]$, which does not. Across 13 coins, the coin-clustered interval is the conservative one, and the honest reading is that the per-fill advantage is not established.

Second, the headline result inverts above the crossover rebate this section measures. The number that shows the arm ahead is per fill at zero rebate. Per quoting opportunity, the arm is better when the rebate is no more than 0.75 bp, and the touch is better at 1.0 bp and above, with the crossing at 0.9076 bp. Both statements are true, but only the second describes the maker who actually chooses between the two rules at this venue. The ordering reverses above the crossover rebate this section measures, and whether any particular maker sits above or below it is a property of that maker's fee schedule rather than of this panel. The crossing is not an independent fact either: it lies within roughly 4% of the touch rule's own break-even rebate, because the arm takes only about 9% of the fills, so it is a ratio of small differences, not a separate measurement. The rebate values quoted here are points on the sweep of the rebate axis, not any venue's published tier; the tier serves here only as a reference point on that axis. As in Section 10, it carries the obligation that neither simulated quoter is shown to satisfy it, so it sets an upper bound on what either could earn, not a description of what they would earn.

Third, what was tested is a depth rule, not the model. By the arm's own criterion, its median quote sits off the touch more than half the time and at the deepest level 57% of the time, so this null covers that family of rules, not a broader one. Two scope notes apply: the window-retention threshold was set to a minimum of 50 fills, rather than the stricter 400 that the same code defaults to, which admits thin windows; and the fill-count floor binds only the arm, not the touch, so the 49 of 140 windows it removes are precisely the ones where the arm quotes deepest, making the attrition informative rather than random.

A note on the per-opportunity metric is warranted, because omitting it would flatter the arm. Multiplying markout by fill rate on a common denominator rewards not trading: a quoter who posts far from the mid does not fill, does not lose, and wins on that metric while being useless. The arm still loses on every fill it takes; it simply takes far fewer.

Measured per opportunity, a quoter posted far from the mid takes almost no fills and so loses almost nothing, which is why the per-opportunity ordering must be read against a rebate paid per fill rather than on its own. The rebate scan is the decision-relevant comparison because rebates are paid per fill.

12 What the Measurements Decide

This section assembles the measurements into the decisions they support for a quoter. It introduces no new measurement; every figure is read from a shipped result file and already appears elsewhere in the paper, with the survivorship bound quoted here derived in Section 14. Because the samples behind the numbers differ, each figure is stated together with the sample it comes from, and the reach of each decision is bounded by that sample.

The entry leg defines the gap that must be overcome. A queue-end quoter at the touch must see the entry markup surpassed by the exit leg and any rebate combined. This threshold equals 0.8387 bp across 3,618 Bybit coin-days, 0.3912 bp across 4,785 Binance USD-M coin-days, and 0.4575 bp across 1,580 Hyperliquid coin-days. A round trip breaks even before fees only if the exit leg and any rebate together cover that threshold. Section 5 gives the accounting identity; the venues' published maker fees appear in Table 1. The Bybit figure carries a survivorship bias of 0.1263 bp coin-weighted and 0.0720 bp fill-weighted under the pre-registered measurement of Section 14, a bias that understates the gap rather than flattering it, so a quoter should treat it as a lower bound. Whether a rebate replaces that fee or is paid on top of it is left to Section 10 and not folded into these three numbers, because that distinction is shown there to shift the crossing one for one.

Bybit's own numbers illustrate the sample dependence. On the buyer-depth subsample of 300 Bybit coin-days, which is 8.3% of that venue's panel and tilted toward 2023, three regions divide the choice among five price rungs from the mid: below 0.3283 bp no rung is profitable per quoting opportunity; between 0.3283 and 0.7711 bp the deepest of the five rungs is preferred; above 0.7711 bp the touch is preferred, because rebates pay per fill and the touch earns far more fills. The touch's own break-even rebate on that subsample is 0.7171 bp, not the 0.8387 bp entry gap above, since the two come from different subsamples, different quoting rules, and fill counts in this paper are indexed to the analysis they belong to (Section 3). The crossover intervals are by-coin [0.7163, 0.8341] and coin-day [0.6753, 0.8698]. Yet the sensitivity that decides the rule sits outside both, because a residual maker fee charged on top of the rebate moves the crossover one for one, so the ordering a maker faces is decided by that maker's own fee schedule rather than by this panel. This makes the rule usable only with that qualification attached.

The crossover sits at 0.7711 bp, so a maker facing a rebate below that figure prefers the deepest rung and one facing a rebate above it prefers the touch; which side a given maker falls on is decided by that maker's own fee schedule rather than by this panel. Any rebate schedule a maker actually faces, however, carries monthly uptime, instrument coverage, and minimum order size obligations. The single-rung, single-instrument, buy-side-only simulated quoter is not shown to satisfy them. It is an upper bound on the rebate, not a rebate the quoter is proven to receive. A separate crossover orders the quoting arm of Section 11 the same way, on a different subsample and between a different pair of rules: across 91 Bybit coin-days, the tested depth rule is preferred at rebates up to 0.75 bp, the touch is preferred at 1.0 bp and above, and the crossover sits at 0.9076 bp.

Measuring the trade sign changes the monitored value, not its sign. For desks without

an exchange aggressor flag that monitor markout themselves, three operating rules follow. Both classifiers keep the adverse-selection drift larger in magnitude than the captured half-spread at all three venues, and the pooled net is negative, so the classifier-based markout keeps its sign. The magnitude moves by less than a quarter of the effect. The direction is not consistent across venues: +3.8% and +3.8% (Bybit), +0.6% and +13.4% (Binance) at the centralized venues, and −20.4% and −11.5% (Hyperliquid) on-chain. No usable constant correction exists, and carrying the centralized adjustment to the on-chain panel would be an error. Resorting to a more accurate rule does not fix this, because agreement rates do not rank the bias across venues: Hyperliquid’s Lee–Ready agreement of 93.6%, against Bybit’s 93.2, does not capture the error, which is 20.4% of Hyperliquid’s own effect versus 3.8% of Bybit’s, about five times farther from zero. That is the reading a desk is most likely to get wrong unaided and the one that travels outside crypto. All three rules carry two qualifications. The point estimates are order-stable but unevenly significant: three of the six paired intervals exclude zero and three contain zero. Part of each error comes from which fills occur rather than how they are signed, with fill counts at 0.930 to 1.082 times the truth. Reading a correction factor from Section 6 corrects both at once.

Across the venues, the panel sometimes cannot support a verdict. When the Kish effective cluster count [22] falls below five, the shipped files read no interval and return no verdict at all, which on this paper’s own result surface withholds 26 of 96 per-coin cells, both horizons for 11 of 20 Hyperliquid coins, the per-coin verdict on the venue difference of Section 8, and the Binance survivorship contrast of Section 14, whose four delisted names fall below the floor on the coin dimension. That floor is not derived; it is the point below which the t critical value moves so quickly that the interval becomes a statement about the critical value rather than about the data, at 3.18 for four effective clusters, 2.78 for five, and 2.31 for nine. A lower floor would return verdicts the estimator cannot support, whereas a higher floor would abstain where the Section 7.2 scan shows the family holds. The shipped reference implementation (Section 13) applies the same rule to order books and flows outside this study, and that rule is the least crypto-specific part of this section.

Nothing above states that any configuration is profitable, because the exit leg, inventory, and all other revenue items are unmeasured here; Section 14 takes up what can still be concluded from these results.

13 Reproducibility

Reproducibility here is claimed narrowly and precisely, and the narrowness is the content. What can be rebuilt is the full chain from frozen derived measurement artifacts upward. All ten aggregate panels reconstruct byte-for-byte from those artifacts. The manifest `reproduce/lineage.json` records every source path relative to the external artifact root, its SHA-256 hash, the hash of each output, and each row count. A source mismatch is rejected before extraction, and every rebuilt output is hash-checked afterward. Appendix B lists the panels and their row counts, and the analysis generator programs rebuild the result files, tables, and figures from those panels.

One further limit sits inside the chain rather than at its edge. The depth-10 archive behind the Bybit panel carries the venue publish stamp `ts`s, and the order-book parser was later changed to prefer `cts` where the archive has it, so a fresh parse of a post-2024-01

Bybit coin-day no longer reproduces the shipped panel. The panel's own basis is retained and recorded per row, because measuring on `cts` would put a book-time shift of 3 ms median on BTC, and 16 ms on a thinner symbol, inside the dead-minus-live contrast. A publish-lagged reference mid is the staleness Proposition 1 says inflates the capture leg, so the choice is recorded rather than silently taken.

The chain terminates at the raw exchange archives and fill-level ledgers, which sit outside the repository for reasons of size and data licensing. The venue-specific measurement programs that convert raw archives into frozen artifacts are not shipped with the package. This repository therefore does not reproduce the study from raw exchange archives, and no claim anywhere in the package asserts otherwise. All three venues share the same outer boundary at the raw archives. However, the depth to which each is reproducible differs: Binance is the shallowest because no fill ledger exists for that venue at all, so its chain starts from derived artifacts.

Shipped alongside the panels, `makercex/fills.py` documents and tests the maker-and-taker rule; it does not reconstruct the raw venue processing pipeline. It now reads the reference mid strictly before each fill, ties broken the same way the measurement path breaks them, so the two agree on the convention that matters: an at-or-before read returns the post-fill book wherever a fill and the update it triggers share a timestamp, which is 52.7% of fills on millisecond-stamped Binance USD-M and drops the quote rule there from 0.996 to 0.620. What the reference implementation still does not do is reconstruct the raw venue processing pipeline or the queue simulator that produced the fills; it documents and tests the fill rule alone.

The automated shipped-versus-rebuilt comparison reaches different depths in different files. In five of them it covers the measurements and not every block beside them: 1,111 of 1,415 leaf nodes in the decomposition file, 130 of 288 in the venue-difference file, 39 of 154 in the depth interval, 68 of 6,575 in the cross-venue lead-lag, and 53 of 1,626 in the quoting arm. A clean comparison there endorses only the numbers it covers, not the whole file. What the comparison does not reach in those five is stored metadata and per-row diagnostic blocks rather than measurements. Three files compare on every leaf they carry and rebuild byte-for-byte: the trade-sign-rule counterfactual on all 510, the per-coin multiplicity file on all 1,164, and the pre-registered survivorship bound on all 567. The per-coin interval file behind Table 10 reaches 1,717 of 1,784. Separately from that comparison, `build_panels.py --verify-shipped` hash-checks every shipped result file against the manifest as well as every panel, so a result file that no longer matches the hash recorded for it fails the package's own check.

Under the package's machine-readable interpretive boundary, `reproduce/claim_contract.json` lists the claims the evidence permits and the claims it forbids, whereas `reproduce/analysis/readme_claims.py` verifies the package's own numeric statements against the shipped result files and rejects forbidden wording.

That contract is the honest remainder after the rebuild chain is stated, and it is where a reader should look before treating any shipped number as settled. The three caveats above are the ones most papers omit, and each is a boundary rather than a defect, which the claim contract shipped with the package records rather than relaxes.

14 Discussion and Limitations

Under the queue-at-the-touch simulation, the central result of this paper is that a simulated maker quoting at the end of the queue at the touch earns a negative pooled 10-second pre-fee entry markout on Bybit, Binance USD-M and Hyperliquid, with two-way clustering on month and coin excluding zero at all three venues. What the preceding sections add is a measured cost of the active-side sign inference on which that result depends, and a boundary on what the same panels can and cannot support. This section states what the result is not, ranks the remaining threats, converts the survivorship-bias argument [6] into a pre-registered measurement, restates the falsification condition, and identifies the two experiments most likely to extend the finding.

The fixed-quoter measurement yields no statement about realized maker profitability. Entry markout does not charge exit costs, does not model inventory, and excludes rebates and all other revenue items. It therefore cannot establish whether market makers at these venues earn or lose money. This paper makes no such claim. It is not a tradable edge, nor a walk-forward strategy result: there is no signal screening, no portfolio construction, and no out-of-sample evaluation, because the fixed-quoter measurement has nothing to screen. It is not a causal comparison of venue architectures, nor a finding of informational leadership. And it is not a claim that a real market maker holding genuine queue priority would observe the same numbers, because the rule in Section 4 deliberately assigns the simulated quoter the worst available queue position.

Eight threats remain, in order. First, weight concentration. The calibration in Section 7.2 shows that when the heaviest clusters are also the atypical ones, the t and two-way estimators run at a size of 0.183 and the guard returns a verdict on only 14% of replications. Bybit BTC is 34.5% of that venue's fills on a capture of 0.0124 bp, which is that shape, and the effective coin counts of 5.96, 10.73 and 5.43 in Section 7.1 are the same concentration seen from the other side. This is the condition these panels demonstrably inhabit, and no diagnostic shipped with the package flags it. The right-skew condition the grid also reports is not the binding one: the grid varies the sigma of a lognormal rather than a skewness, and the most right-skewed distribution these panels carry, Hyperliquid's across calendar months at +2.02 against -0.68 on Bybit and +0.45 on Binance, maps to the grid's skewness-2.0 row, where every method sits near nominal at 0.057 to 0.100. The size of 0.42 belongs to an implied skewness of 414. Second, the Hyperliquid panel is fragile on every dimension simultaneously: its weighting spread exceeds its own smallest estimate, its month skew is +2.02, 11 of 20 per-coin verdicts are withheld, 7 coins carry positive point estimates, and its 60-second two-way interval cannot be constructed. Its pooled verdict excludes zero, and it is the weakest of the three. Its pooled figure is also the most weighting-dependent of the three: four coins carry 80% of its fills, and weighting coins equally instead moves the pooled net from -0.4575 to -0.1470 bp, at which a one-sample t on the twenty coin means gives $t = -2.10$ and the coin sign count is 13 of 20. The abstract's claim that the interval excludes zero at all three venues rests on the fill weighting at this venue. To all of that add the extraction discrepancy of Section 8: the two shipped extractions of the same 500 coin-days, run at the same 100 ms re-quote ceiling, carry fill sets differing by a factor of two and nets of -0.5825 against -0.3435 bp, where Bybit's two extractions agree fill for fill. The headline figure for this venue is read from the main panel, so one of the two extractions is wrong and the size of the error is of the same order as the estimate. Of everything listed here this is the item that would most change the paper if it resolved against the panel, and settling it needs the venue-specific measurement

programs that Section 13 places outside the package. Third, the panels are not comparable: calendars, coin sets and sampling cadences differ, so the three rows of Table 3 are three independent measurements, not a comparison. Fourth, the split between adverse selection and capture depends on sampling cadence in the manner made explicit by Proposition 1 and Table 2, so every share statement in this paper inherits that dependence. Fifth, the panels are not point-in-time survivorship-aware panels, treated in the next paragraph as a decision rather than an open threat. Sixth, the CME contrast in the abstract and Figure 1 is confounded. The companion paper places its order at the front of the remaining queue, while this one places it at the tail, so the two panels differ in queue rule as well as asset class, period and data source. The inversion is not attributable to the asset class until a front-of-queue arm is run on the same panels. Seventh, the fill rule conditions on the aggressor’s size exceeding the queue ahead of the order. This selects the large book-clearing trades that carry the highest expected post-fill drift. Part of the measured negative markout could therefore be a consequence of that selection rather than of quoting at the touch, and the shipped panels bound how much. Each panel records a mean fill size per coin-day as a fraction of the posted one-coin order, so sorting each coin’s days into terciles of that fraction and taking the fill-weighted net in each tercile tests whether the days with the largest fills carry the result. They do not, but the proxy is censored: realised fill size is capped at the posted one-coin order. The panels sit within a thousandth of that cap on 25% of Bybit coin-days, 50% of Binance and 32% of Hyperliquid, and exactly at it on 24, 41 and 31%. The top tercile is largely a tie at the cap, and the split separates only the low end. Bybit gives -0.8318 , -0.8566 and -0.8267 bp across the three terciles, which is flat; Binance gives -0.4277 , -0.3673 and -0.3749 bp, so its smallest fills are the most negative, the opposite of the selection story; only Hyperliquid is ordered in the direction of concern, at -0.4219 , -0.4456 and -0.5156 bp, a spread of 0.094 bp against a headline of -0.4575 bp. Mean fill size is also a coin-day average rather than a per-fill quantity. The test therefore bounds the part of the selection channel visible below the cap rather than the channel as a whole, and the quantity that would bound it fully, the aggressor’s trade size against displayed queue depth, is not in the shipped panels.

Eighth, the entire dataset is simulated fills from a queue simulator that is never validated against an observed maker fill on any venue. The shipped reference implementation documents and tests the fill rule and now applies the measurement’s own tie-break, but it is not the simulator: no shipped code produces the fills, so the instrument itself remains outside every external check. Hyperliquid’s order-level feed would make such a validation feasible, but it is not attempted here. What exists instead is an external comparison rather than a validation: Albers and coauthors [5] measure this paper’s estimand on observed Hyperliquid fills and report roughly -0.48 bp before rebate for the average maker fill, against -0.4575 bp here for the venue and -0.5832 bp for its BTC, the instrument they measure. That the simulator lands where observed fills land is reassuring about the level it produces on one venue at one horizon. It is not a test of the simulator, because their fills are not the fills it generates and no fill-by-fill correspondence is checked.

No headline panel in this paper contains delisted coins, and the direction of that omission was previously an argument from mechanism, not a measurement of it. Reading each venue’s own archives supplies the missing cross-section, the names that died within each window: 30 of the 206 perpetuals Bybit listed at the start of its window and 4 of Binance’s 185, counts drawn from the raw archives that Section 13 places outside the shipped package and that therefore cannot be rebuilt from it. The ten longest-surviving Bybit delistings and all four qualifying Binance instruments were re-run through

the same decomposition (3,451 coin-days, shipped as `reproduce/panels/delisted_coindays.csv` and analysed in `reproduce/survivorship_bound.json`). The prediction, coin set, matching-date rule and two weighting schemes were fixed in advance in `reproduce/preregistration/survivorship_bound.md`. Its SHA-256 is recorded in `reproduce/preregistration/LOCK_HASHES.md` as `a53b31606661adc36658f09f3d10ab5daaa6caef31236b32bcbb6abb064081e5`. That commit contains no line of the delisted panel, which is consistent with the plan preceding the measurement. The ordering is nonetheless recorded rather than independently verifiable because the repository and its history are author-controlled. A reader who wants an independent guarantee would need a timestamp this package cannot provide. (A later sibling file carries wording revisions and its own digest; the file named above is the one that predates the measurement.)

On matched dates, the delisted names underperform survivors by 0.8483 bp, with a two-way month-and-coin interval of $[-1.0532, -0.6433]$ and all ten negative. Averaging over the ten coins rather than over fills gives a gap of 0.8673 bp; multiplying this coin-weighted gap by the 14.6% death rate (which is the rate for Bybit's full listed universe, not for our panel of 13 large liquid survivors, so this correction is an upper bound on the omission rather than an estimate of it, an assumption we flag rather than a fact) yields an omission of 0.1263 bp coin-weighted, whereas multiplying the fill-weighted gap of 0.8483 bp by the dead names' 8.5% share of fills in the measured set yields an omission of 0.0720 bp fill-weighted; the interval endpoint is not carried into a bound because the two weightings cannot be mixed. All in the direction of making the measured net less negative than a survivorship-free panel would show. The fill-weighted number is smaller because the main result's weights concentrate on the largest names, which are precisely the least affected by the omission. Three qualifications bound how far this conclusion travels. First, the delisted names trade thinly (1,629 fills per day against 12,569 for the surviving panel), so the comparison prices thinness along with death. Benchmarking the same coin-days against only the least active third of the surviving panel narrows the activity ratio from 7.7 to 2.8 and leaves a gap of 0.6304 bp. Second, the ten are the longest-surviving qualifying delistings, and therefore the least distressed of the delisted names that qualified, so the gap is more likely understated than overstated. Third, the evidence supporting it is consistency across ten coins and 28 months, not the interval, because the monthly means are sign-consistent, so month-resampling cannot reach zero.

The other two venues neither confirm nor contradict it. Binance USD-M: the same test gives +0.0622 bp with an interval of $[-0.9631, +1.0875]$, and its verdict is withheld rather than reported as a failure to clear zero, because its 319-day window contains only four delistings, two of which are contracts on index products rather than coins, and four coins sit below the effective-cluster floor that governs every other verdict here. Dropping the two index contracts leaves +0.0320 bp on the remaining pair, withheld on the same ground. What binds this venue is in any case not the gap but the death rate: at 4 of 185 names, the omission cannot move its pooled number by more than 0.0235 bp even at the far end of the interval. Hyperliquid is untested because both of its archive buckets are requester-pays, and no listing or delisting record is obtainable. In sum: the survivorship direction is now measured on one of three venues, unresolved on the second, and open on the third, and none of this makes any panel point-in-time. A replication that wants that still needs what it always needed: point-in-time listing histories for each venue and window, the inclusion rules behind coin lists and cadences, and the same decomposition re-run over the implied universe.

Falsification would come from a venue, coin or state cell whose adverse selection

reliably falls below the captured half-spread, with net excluding zero on the positive side. The bar differs by the level the cell sits at, because the two levels do not carry the same clustering dimensions, and stating one bar for both would set a test no coin-level cell could ever meet. A pooled venue cell must survive what the three negative pooled cells survived: same sign under fill-weighting and coin-weighting, holding at both the 10-second and 60-second horizons, with the interval clustered on month and coin simultaneously and an effective cluster count above the abstention threshold on both dimensions. A per-coin cell carries only the month dimension, as Section 7.3 records, so the bar it must meet is the one Table 10's negative cells actually met and no more: a month-clustered interval excluding zero on the positive side at both horizons, on an effective month count above the same threshold. That is the whole of it. Of the 56 negative cells that clear, 49 also carry monthly means of one sign, so that consistency across months rather than the interval is what holds them up, but seven do not, and a positive cell is not asked to supply what those seven did not. A per-coin counterexample clearing this bar would not by itself carry a two-way verdict, and none is demanded of it. This is not hypothetical: seven of Hyperliquid's twenty coins carry positive 10-second point estimates, led by NEAR at +0.4029 bp. However, five of those seven carry a withheld verdict, so they were not tested at all; only two were tested, and neither excludes zero. Across all three venues, the shipped panels record no coin significantly positive, and no counterexample overturning non-positive net. A replication producing such a cell at any venue would overturn the claim, and the shipped per-coin files are arranged so that such a cell is immediately visible.

Two levers point where to push next. First, the measurement rule is the largest open lever: a quoter holding genuine time priority, replacing the one placed at the end of the queue at every snapshot, would move the capture leg up by an amount this paper cannot itself bound, and it is the single change most likely to flip the sign at the venues where net is smallest. External evidence now says how large that move might be: on Hyperliquid the gap between the average maker fill and a first-in-line cohort is about 0.8 bp before rebate [5], which exceeds two of the three nets reported here and is close to the third. A front-of-queue arm run on these panels is therefore not a robustness check but the experiment that decides whether the sign is a property of quoting at the touch or of quoting at the back of it. Second, a depth frontier that accounts for residual fees, because Section 10 shows that dimension dominates the interval currently reported alongside it. Neither is a patch to what exists; both are experiments that would decide how far the result travels.

15 Conclusion

The 10-second pre-fee entry markout is negative on all three venues: -0.8387 bp on Bybit across 3,618 coin-days, -0.3912 bp on Binance USD-M across 4,785, and -0.4575 bp on Hyperliquid across 1,580, with the two-way month-and-coin interval excluding zero on every exchange. This stands in contrast to the near-exact cancellation measured on CME futures in the companion paper, where the captured half-spread and the adverse-selection drift are almost equal and the net is indistinguishable from zero, although this contrast is confounded because the companion paper places its order at the front of the remaining queue while this one places it at the tail, so the inversion is not attributable to the asset class until a front-of-queue arm is run.

What survives is the sign. The result holds under four units, the two weightings crossed with both markout horizons, dropping any single coin on any exchange, and

replacing the exchange aggressor flag with trade-sign classifiers on all three venues. The two 60-second Hyperliquid units are withheld from this count, as are the three pooled venue nets. Those pooled nets are the headline result rather than a perturbation the sign survives. The magnitude in its neighbourhood is a weaker object, ranked in Section 14, headed by the condition the calibration identifies and these panels demonstrably inhabit: when the heaviest clusters are also the atypical ones, every interval method here runs at a size of 0.183 and returns a verdict on only 14% of replications. The finding stands, and the limits under which it stands are stated. Measurements of this kind are typically reported under a single clustering choice, with alternatives left uncomputed; this paper computed the alternatives, showed where they diverge, and located the region where the whole family fails together. The result extends only as far as the pessimistic queue assumption and the declared data boundary permit, and no further.

A Per-Coin

Table 10 reports the full per-coin surface behind Figure 2, fill-weighted at the 10-second horizon in bp with 95% month-clustered t intervals. The verdict column records “clears” where the interval excludes zero, “no” where it does not, and “withheld” where the Kish effective month count falls below the threshold at which the interval is interpretable. Binance coin names carry the exchange’s USDT suffix in the packaged files and are stripped in the table.

Table 10: Per-coin capture, adverse selection and net entry markout, fill-weighted, in basis points at the 10-second horizon, with 95% month-clustered t intervals and the verdict each interval supports. Withheld marks a cell whose interval is not worth reading. Source: `reproduce/decomposition_by_venue.json`, `reproduce/per_coin_intervals.json`.

Venue	Coin	Capture	Adverse Selection	Net	95% CI on net	adv/cap	Trades	Verdict
Bybit	ARB	0.5937	-1.7440	-1.1503	[-1.287, -1.013]	2.94	1,276,082	clears
Bybit	AVAX	0.5687	-1.5775	-1.0088	[-1.074, -0.943]	2.77	1,473,115	clears
Bybit	BNB	0.5381	-1.4531	-0.9150	[-1.001, -0.828]	2.70	1,561,877	clears
Bybit	BTC	0.0124	-0.6218	-0.6094	[-0.644, -0.574]	50.17	15,687,748	clears
Bybit	DOGE	0.3282	-1.2234	-0.8952	[-0.975, -0.814]	3.73	3,461,868	clears
Bybit	ETH	0.0237	-0.8586	-0.8349	[-0.883, -0.786]	36.26	6,289,205	clears
Bybit	HYPE	0.3906	-2.0832	-1.6927	[-1.988, -1.397]	5.33	1,363,519	clears
Bybit	LINK	0.3794	-1.4122	-1.0329	[-1.114, -0.951]	3.72	2,202,521	clears
Bybit	LTC	0.5884	-1.6741	-1.0857	[-1.203, -0.967]	2.85	1,157,268	clears
Bybit	OP	0.3887	-1.5367	-1.1480	[-1.272, -1.023]	3.95	1,769,720	clears
Bybit	SOL	0.2439	-1.0935	-0.8497	[-0.901, -0.797]	4.48	4,636,848	clears
Bybit	SUI	0.3228	-1.3138	-0.9910	[-1.089, -0.892]	4.07	2,389,646	clears
Bybit	XRP	0.3622	-1.1191	-0.7569	[-0.784, -0.729]	3.09	2,206,601	clears
Binance	1000PEPE	0.7723	-1.0997	-0.3274	[-0.384, -0.269]	1.42	9,036,389	clears
Binance	1000SHIB	0.6981	-1.1248	-0.4267	[-0.537, -0.315]	1.61	4,109,443	withheld
Binance	ADA	1.2524	-1.4261	-0.1737	[-0.286, -0.060]	1.14	1,569,964	clears
Binance	ARB	0.6218	-1.0914	-0.4696	[-0.526, -0.412]	1.76	3,470,244	clears
Binance	AVAX	0.5113	-0.9295	-0.4182	[-0.501, -0.334]	1.82	4,735,762	clears
Binance	BNB	0.3008	-0.6943	-0.3935	[-0.433, -0.353]	2.31	3,686,802	clears
Binance	BTC	0.0352	-0.3597	-0.3245	[-0.363, -0.285]	10.21	13,729,971	clears
Binance	DOGE	0.7233	-1.1031	-0.3798	[-0.463, -0.296]	1.53	3,371,580	withheld
Binance	ETH	0.0509	-0.5847	-0.5338	[-0.595, -0.472]	11.48	8,312,432	clears
Binance	LINK	0.6124	-1.0552	-0.4428	[-0.518, -0.367]	1.72	3,370,020	clears
Binance	LTC	0.8023	-1.1318	-0.3295	[-0.421, -0.237]	1.41	1,835,855	clears
Binance	OP	0.5565	-1.0398	-0.4833	[-0.526, -0.440]	1.87	5,427,820	clears
Binance	SOL	0.3153	-0.7203	-0.4050	[-0.465, -0.344]	2.28	9,424,792	clears
Binance	SUI	0.8410	-1.2504	-0.4093	[-0.477, -0.341]	1.49	4,155,295	clears
Binance	XRP	1.0272	-1.1259	-0.0987	[-0.202, +0.004]	1.10	2,172,255	no
Hyperliquid	AAVE	1.2928	-1.6420	-0.3491	[-0.981, +0.282]	1.27	30,596	withheld
Hyperliquid	APT	1.6161	-1.5546	+0.0615	[-0.180, +0.303]	0.96	12,978	no

Venue	Coin	Capture	Adverse Selection	Net	95% CI on net	adv/cap	Trades	Verdict
Hyperliquid	ARB	1.4602	-1.1942	+0.2660	[-1.288, +1.820]	0.82	16,361	withheld
Hyperliquid	AVAX	1.2923	-1.3182	-0.0258	[-0.928, +0.876]	1.02	22,398	no
Hyperliquid	BNB	0.8045	-1.0173	-0.2128	[-0.968, +0.543]	1.26	28,555	no
Hyperliquid	BTC	0.1819	-0.7651	-0.5832	[-0.707, -0.459]	4.21	541,801	clears
Hyperliquid	DOGE	0.8488	-0.9046	-0.0557	[-0.960, +0.849]	1.07	47,077	withheld
Hyperliquid	ETH	0.3283	-1.0293	-0.7010	[-0.796, -0.605]	3.14	352,945	withheld
Hyperliquid	HYPE	0.6810	-1.0520	-0.3710	[-0.753, +0.011]	1.54	483,387	no
Hyperliquid	INJ	2.5992	-2.7078	-0.1086	[-0.956, +0.739]	1.04	7,983	withheld
Hyperliquid	LINK	1.3968	-1.2403	+0.1565	[-1.735, +2.048]	0.89	36,984	withheld
Hyperliquid	LTC	1.6206	-1.6015	+0.0191	[-1.320, +1.358]	0.99	16,425	withheld
Hyperliquid	NEAR	2.1640	-1.7612	+0.4029	[-0.118, +0.924]	0.81	15,813	no
Hyperliquid	OP	2.1830	-1.9034	+0.2796	[-0.874, +1.433]	0.87	7,656	withheld
Hyperliquid	SEI	1.8221	-1.7016	+0.1205	[-0.266, +0.507]	0.93	9,986	withheld
Hyperliquid	SOL	0.4457	-1.0553	-0.6096	[-0.808, -0.410]	2.37	173,259	clears
Hyperliquid	SUI	0.8782	-1.3170	-0.4388	[-0.673, -0.203]	1.50	40,019	withheld
Hyperliquid	TIA	1.4151	-1.7549	-0.3399	[-1.371, +0.692]	1.24	10,121	no
Hyperliquid	WLD	1.9960	-2.2191	-0.2231	[-1.925, +1.478]	1.11	17,482	no
Hyperliquid	XRP	0.5361	-0.7640	-0.2279	[-0.927, +0.472]	1.43	72,454	withheld

B Panel Lineage

Each panel is rebuilt byte-for-byte from a frozen derived artifact whose SHA-256 hash is recorded in `reproduce/lineage.json`, and every rebuilt output is hash-checked against the shipped copy. The frozen artifacts, the underlying exchange archives, and the exchange-specific measurement programs connecting them lie outside the package. Table 11 lists the ten shipped aggregate panels, their row counts, and the analyses they carry.

Table 11: The ten shipped aggregate panels

Panel	Rows	Carries
bybit_perp_ coindays.csv	3,618	Bybit decomposition and per-coin intervals
binance_um_ coindays.csv	4,785	Binance decomposition and per-coin intervals
hyperliquid_ coindays.csv	1,580	Hyperliquid decomposition and per-coin intervals
bybit_depth_levels. csv	1,500	Depth and rebate frontier, with intervals
bybit_quoting_arm_ coindays.csv	91	Quoting arm
cross_venue_ leadlag_coindays. csv	325	Lead-lag
bybit_conditional_ cells.csv	7,026	Conditional decomposition, session and regime
venue_requote_ rungs.csv	5,775	Venue difference and re-quote rung sweep
sign_rule_coindays. csv	29,949	Trade-sign counterfactual
delisted_coindays. csv	3,451	Survivorship bound for delisted perpetuals

The ten aggregate panels shipped with the paper, each with its row count and the analyses it carries. Source: `reproduce/lineage.json`.

C Figure Provenance

Table 12: Figure provenance

Figure	Source
Fig. 1 pooled decomposition	reproduce/decomposition_by_venue.json; where the CME reference group is hardcoded from the accompanying [18], and is not produced by any panel here
Fig. 2 per-coin surface	reproduce/decomposition_by_venue.json, reproduce/per_coin_intervals.json
Fig. 3 sign-rule cost	reproduce/sign_rule_counterfactual.json
Fig. 4 conditional contrast	reproduce/conditional_net_asia.json and the corresponding files for the European and US sessions
Fig. 5 depth and rebate frontier	reproduce/depth_rebate_frontier.json

Each figure and the shipped result file its numbers are rendered from. All five regenerate from reproduce/make_figures.py; the CME reference group in Figure 1 is carried over from [18] and is not produced by any panel here.

D Results Inventory

Table 13: Every analysis in Sections 5 to 13 with the verdict it returned, so that the withheld and null results sit in the same list as the central one rather than behind it. Each row names its own source file where one is quoted in the body.

analysis	verdict
decomposition (central)	The pooled 10-second pre-fee net entry markout is negative on all three venues (-0.8387 , -0.3912 , -0.4575 bp), with month and coin two-way intervals excluding zero; stable under two weightings, two horizons and one-by-one coin exclusion, with the two 60-second Hyperliquid units withheld
per coin	Bybit 13 of 13 negative with intervals excluding zero; Binance 15 of 15 negative, 12 excluding zero, 1 not excluding, 2 not given a verdict; Hyperliquid 13 of 20 negative at 10 seconds, 2 excluding zero, 11 not given a verdict. Of 96 units, 57 have intervals that cannot touch zero; 26 give no verdict
the ratio	Pooled values 1.8 to 5.1, with corresponding coin medians 3.7, 1.7, and 1.1; the denominator collapses on the highest-weighted coins. Not a core effect
sign-rule counterfactual	Lee-Ready and tick rule give the same direction as the venue for 93.2% and 89.8% of Bybit trades, 98.2% and 91.3% of Binance, and 93.6% and 89.4% of Hyperliquid, respectively. The sign of the net amount survives under both rules on all three venues; the magnitude changes by +3.8% and +3.8% on Bybit, +0.6% and +13.4% on Binance, and -20.4% and -11.5% on Hyperliquid, with three of six intervals excluding zero. Agreement rates do not give a cross-venue ranking of errors
inference	Relative to month-only, two-way clustering inflates standard errors by 6.0, 2.0, and 1.2 times, respectively; all verdicts survive

analysis	verdict
interval calibration	In this paper’s own clustering, a coin effect persisting across months takes the month-only method to a test level of 0.867 while the two-way method holds at 0.057, and a shock shared by every coin on one date leaves both at 0.052. In the coin-clustered grid beside it, weight concentration breaks all four methods together at 0.183, with a verdict returned on 14% of replications. Source: reproduce/coverage_sweep.json
sign-flip calibration	On each venue’s own months, test levels are 0.053, 0.061, and 0.124, with calibration recorded in reproduce/wild_test_calibration.json; no packaged verdict uses this test
venue difference	Same-day Bybit minus Hyperliquid pooled value -0.3546 bp, negative on 13 of 13 coins and 293 of 325 coin-days; depends on extraction method (-0.3496 vs -0.2232), rung fixed, and both clustering verdicts are not given. Descriptive, not causal
lead-lag	On 322 of 325 coin-days, Bybit leads by a median of 500 ms, corresponding to a 441 ms release cadence difference. Cannot distinguish leading from release frequency differences; no verdict given
conditional	Positive on 7, 7, and 9 of 12 coins, respectively; $p = 0.789, 0.772, 0.285$. This is a failure to reject under low power, not invariance
depth and rebate	Every rung is negative before rebates; the minimum break-even rebate is 0.328 bp, and the best bid/offer exceeds the deepest rung at 0.771 bp, coin-clustered [0.716, 0.834] and coin-day [0.675, 0.870]; a residual maker fee charged on top of the rebate shifts the crossing one for one, a dimension the bootstrap does not cover
quoting arm	Per-trade performance is better (-0.5439 vs -0.9098 bp), but negative on 66 of 91 coin-days and only 9.2% of trades; not significant under coin clustering; above the 0.9076 bp crossing the preference reverses to the best bid/offer
reproducibility	All ten panels rebuild byte-for-byte under hash verification; raw archives and trade ledgers are outside the boundary

Data and Code Availability

The reproducible package is available at <https://github.com/DaruFinance/crypto-adverse-selection> under the MIT licence. It contains a reference implementation of the maker-and-taker fill rule, which documents that rule but does not reproduce the measurement path’s tie-breaking and is not the program that produced the fills, the ten aggregated coin-day panels, the analysis programs that generate every result, the accompanying result files, the figure-generation programs, and a machine-readable claims contract. It does not contain any raw exchange data; the aggregated panels are provided for reproducibility subject to the originating exchanges’ data terms. Section 13 states precisely where the reproducible chain begins and where it ends.

References

- [1] Albers, J., Cucuringu, M., Howison, S., & Shestopaloff, A. Y. (2025). The Good, the Bad, and Latency: Exploratory Trading on Bybit and Binance. *Quantitative Finance* 25(6), 919–947.
- [2] Albers, J., Cucuringu, M., Howison, S., & Shestopaloff, A. Y. (2025). The Market Maker’s Dilemma: Navigating the Fill Probability vs. Post-Fill Returns Trade-Off. arXiv:2502.18625. <https://arxiv.org/abs/2502.18625>
- [3] Albers, J., Cucuringu, M., Howison, S., & Shestopaloff, A. Y. (2026). The Price Impact of Nothing: Rejected Orders as Predictors of Future Returns. SSRN Working Paper 6250378, February 2026. <https://ssrn.com/abstract=6250378>
- [4] Albers, J., Cucuringu, M., Howison, S., & Shestopaloff, A. Y. (2026). An Open Book: Level 4 Order Book Data from the Hyperliquid Exchange. SSRN Working Paper 6465720, March 2026. <https://ssrn.com/abstract=6465720>, doi:10.2139/ssrn.6465720
- [5] Albers, J., Cucuringu, M., Howison, S., & Shestopaloff, A. Y. (2026). The “Neutrinos” of the Order Book: Pervasive, Weakly Interacting Order Flow and its Consequences. SSRN Working Paper 7162966, July 2026. <https://ssrn.com/abstract=7162966>
- [6] Ammann, M., Burdorf, T., Liebi, L., & Stöckl, S. (2022). Survivorship and Delisting Bias in Cryptocurrency Markets. SSRN Working Paper 4287573. <https://ssrn.com/abstract=4287573>
- [7] Andersen, T. G., & Bondarenko, O. (2014). Assessing Measures of Order Flow Toxicity via Perfect Trade Classification. CREATES Research Paper 2014-30, Aarhus University.
- [8] Avellaneda, M., & Stoikov, S. (2008). High-Frequency Trading in a Limit Order Book. *Quantitative Finance* 8(3), 217–224.
- [9] Barone, D., & Lillo, F. (2026). Trading in the Sunshine or in the Shade: Market Impact and Adverse Selection on Hyperliquid. arXiv:2606.15715, June 2026. <https://arxiv.org/abs/2606.15715>
- [10] Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society, Series B* 57(1), 289–300.
- [11] Benjamini, Y., & Yekutieli, D. (2001). The Control of the False Discovery Rate in Multiple Testing Under Dependency. *Annals of Statistics* 29(4), 1165–1188.
- [12] Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2008). Bootstrap-Based Improvements for Inference with Clustered Errors. *Review of Economics and Statistics* 90(3), 414–427.
- [13] Cameron, A. C., Gelbach, J. B., & Miller, D. L. (2011). Robust Inference with Multiway Clustering. *Journal of Business & Economic Statistics* 29(2), 238–249.
- [14] Cartea, Á., Jaimungal, S., & Penalva, J. (2015). *Algorithmic and High-Frequency*

Trading. Cambridge University Press.

- [15] Chakrabarty, B., Pascual, R., & Shkilko, A. (2015). Evaluating Trade Classification Algorithms: Bulk Volume Classification versus the Tick Rule and the Lee–Ready Algorithm. *Journal of Financial Markets* 25, 52–79.
- [16] DeLise, T. (2024). The Negative Drift of a Limit Order Fill. arXiv:2407.16527. <https://arxiv.org/abs/2407.16527>
- [17] Easley, D., López de Prado, M., & O’Hara, M. (2016). Discerning Information from Trade Data. *Journal of Financial Economics* 120(2), 269–285.
- [18] Gatto, D. V. (2026). Adverse Selection Consumes the Touch: A True-Aggressor-Signed Maker-P&L Decomposition Across 13 CME Futures. SSRN Working Paper 7022599, June 2026. <https://ssrn.com/abstract=7022599>, doi:10.2139/ssrn.7022599
- [19] Gatto, D. V. (2026). No Edge Without Information: An Empirical Study of Tradeability in Decentralized-Exchange-Only Cryptocurrencies. SSRN Working Paper 6858778. <https://ssrn.com/abstract=6858778>
- [20] Glosten, L. R., & Milgrom, P. R. (1985). Bid, Ask and Transaction Prices in a Specialist Market with Heterogeneously Informed Traders. *Journal of Financial Economics* 14(1), 71–100.
- [21] Guéant, O., Lehalle, C.-A., & Fernandez-Tapia, J. (2013). Dealing with the Inventory Risk: A Solution to the Market Making Problem. *Mathematics and Financial Economics* 7(4), 477–507.
- [22] Kish, L. (1965). *Survey Sampling*. Wiley, New York.
- [23] Lee, C. M. C., & Ready, M. J. (1991). Inferring Trade Direction from Intraday Data. *Journal of Finance* 46(2), 733–746.
- [24] MacKinnon, J. G., & Webb, M. D. (2017). Wild Bootstrap Inference for Wildly Different Cluster Sizes. *Journal of Applied Econometrics* 32(2), 233–254.
- [25] Millionis, J., Moallemi, C. C., Roughgarden, T., & Zhang, A. L. (2022). Automated Market Making and Loss-Versus-Rebalancing. Working Paper. arXiv:2208.06046. <https://arxiv.org/abs/2208.06046>
- [26] Moallemi, C. C., & Yuan, K. (2016). A Model for Queue Position Valuation in a Limit Order Book. Working Paper, Columbia University.
- [27] Odders-White, E. R. (2000). On the Occurrence and Consequences of Inaccurate Trade Classification. *Journal of Financial Markets* 3(3), 259–286. doi:10.1016/S1386-4181(00)00006-9
- [28] Petersen, M. A. (2009). Estimating Standard Errors in Finance Panel Data Sets: Comparing Approaches. *Review of Financial Studies* 22(1), 435–480.
- [29] Tiniç, M., Şensoy, A., Akyildirim, E., & Corbet, S. (2023). Adverse Selection in Cryptocurrency Markets. *Journal of Financial Research* 46(2), 497–546.